SEMMA

Neel Desai

September 2023

Abstract

This paper into the domain of association rule mining, a powerful data mining technique used to identify relationships and correlations in large datasets. In the context of mobile app usage, this technique can be used to discern patterns of app usage, offering insights into user behavior. For instance, the potential association between a user opening a messaging app right after checking their email is explored. Such insights can enhance user experience through app recommendations or optimization of app placements.

The paper introduces the SEMMA methodology (Sample, Explore, Modify, Model, and Assess) developed by the SAS Institute. This approach provides a structured method for data mining tasks, ensuring the insights derived are robust and interpretable. The primary focus of this research is to harness the power of association rule mining within the SEMMA framework to predict app usage patterns. Using a dataset chronicling individual mobile usage, the research aims to uncover rules that predict the likelihood of opening certain apps after using a specific app, potentially leading to features like "suggested apps" that anticipate a user's next move.

1 Introduction

Association rule mining is a powerful data mining technique that seeks to identify relationships and correlations among items in large datasets. In the realm of mobile app usage, association rule mining can be employed to discern patterns of app usage, offering insights into user behavior. For instance, if a user frequently opens a messaging app shortly after checking their email, there exists a potential association rule between the two apps. Such insights can be instrumental in enhancing user experience by recommending apps or optimizing app placements.

The SEMMA methodology, an acronym for Sample, Explore, Modify, Model, and Assess, offers a systematic approach to data mining tasks. Developed by SAS Institute, SEMMA streamlines the process of extracting meaningful patterns and relationships from large datasets, ensuring that the derived insights are both robust and interpretable.

In this research, we harness the power of association rule mining within the framework of the SEMMA methodology to predict app usage patterns. Using a dataset that chronicles the mobile usage of an individual, we aim to uncover rules that indicate the likelihood of opening certain apps after using a particular app. Such findings can pave the way for features like "suggested apps" that enhance user experience by anticipating their next move.

2 Methodologies

The SEMMA methodology stands as a cornerstone in the realm of data mining. It offers a systematic approach to model-building, ensuring data-driven decisions are sound, replicable, and actionable. Let's delve into the details of each phase:

1. **Sample:** The journey begins by taking a representative subset of the dataset. This serves two purposes: speeding up computations and offering a manageable, yet informative, slice of data.
2. **Explore:** During this phase, a comprehensive understanding of the data is established. This entails visualizations, summary statistics, and identifying potential relationships or anomalies.
3. **Modify:** Preprocessing steps are taken in this phase to prepare the data for modeling. This could involve data transformation, feature engineering, or handling missing values.
4. **Model:** This is the phase where the actual model is built using appropriate algorithms. In our context, this involves association rule mining.
5. **Assess:** Finally, the model's performance is gauged. This involves evaluating the results, ensuring that the model meets the desired objectives, and is fit for deployment.

Incorporating SEMMA into our research ensures a methodical and rigorous approach, allowing for consistent and reliable results.

3 Implementation

Aligning with the SEMMA methodology, the implementation follows a structured approach:

1. **Sample:**
    - The dataset, sourced from Kaggle, consists of mobile usage data for an individual. This includes timestamps for when an app was opened and the duration of its usage.
    - Given that the dataset comprises 4015 entries, there wasn't a necessity to extract a subset. Hence, all records were retained for analysis.

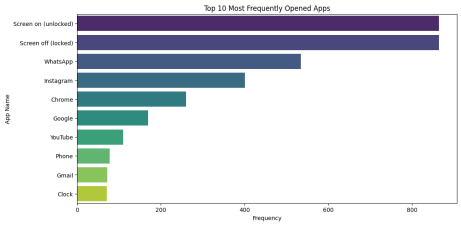

Figure 1: Frequently used apps

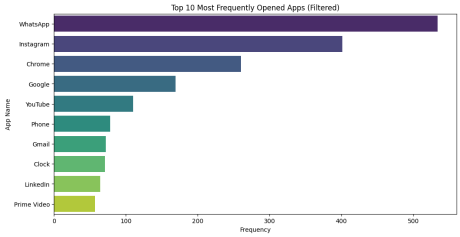

Figure 2: Frequently Used Apps Filtered

- Initial data cleaning was performed, ensuring there were no missing values.

2. **Explore:**

   - An exploratory data analysis (EDA) was conducted to gauge the distribution and patterns present within the dataset.
   - The most frequently opened apps were identified.
   - Preliminary patterns and sequences, which could be vital for rule generation, were investigated.

3. **Modify:**

   - The data was preprocessed to suit the format required for association rule mining.
   - An important insight from the EDA was that events like 'Screen on (unlocked)' and 'Screen off (locked)' were frequent. However, such events are not meaningful for rule generation. Hence, they were excluded from the dataset.
   - The sequences of app usage were prepared for transformation into a transactional format, typically one-hot encoding.
   - The 'Date' and 'Time' columns were combined to form a singular 'Datetime' column, which provides a chronological sequence of app usage.

Out[19]:

	App name	Date	Time	Duration	Datetime	Time_Diff	Transaction_ID
4011	WhatsApp	01/08/23	11:47:31 pm	0:01:08	2023-01-08 23:47:31	NaN	0
4010	Google	01/08/23	11:48:44 pm	0:00:04	2023-01-08 23:48:44	1.216667	0
4009	Chrome	01/08/23	11:48:48 pm	0:00:37	2023-01-08 23:48:48	0.066667	0
4008	Google	01/08/23	11:49:25 pm	0:00:00	2023-01-08 23:49:25	0.616667	0
4007	WhatsApp	01/08/23	11:49:27 pm	0:00:05	2023-01-08 23:49:27	0.033333	0
4006	WhatsApp	01/08/23	11:49:37 pm	0:00:08	2023-01-08 23:49:37	0.166667	0
4003	WhatsApp	01/08/23	11:50:06 pm	0:00:01	2023-01-08 23:50:06	0.483333	0
4002	Instagram	01/08/23	11:50:11 pm	0:01:07	2023-01-08 23:50:11	0.083333	0
4001	WhatsApp	01/08/23	11:51:20 pm	0:08:53	2023-01-08 23:51:20	1.150000	0

Figure 3: Transactions Data

Out[22]:

	Rule	Support	Confidence
6	Google => Chrome	0.091097	0.745763
7	Google => Instagram	0.082816	0.677966
5	Google => WhatsApp	0.080745	0.661017
23	Gmail => Instagram	0.068323	0.611111
10	Chrome => WhatsApp	0.115942	0.583333

Figure 4: Associate Rules

- A 'Transaction_ID' was assigned to groups of app usage events based on a threshold time difference. Any gap greater than 5 minutes between consecutive app openings was considered the start of a new transaction.
- The data was grouped by 'Transaction_ID' to derive a list of apps accessed in each transaction. This data was then one-hot encoded to facilitate the association rule mining process.
- The Apriori algorithm was manually implemented to discover frequent single-item itemsets. The support metric was computed for each app, and those surpassing a minimum threshold (10% in this case) were deemed frequent.

4. **Model:**

- The Apriori algorithm was manually implemented to discover frequent single-item itemsets. The support metric was computed for each app, and those surpassing a minimum threshold (10% in this case) were deemed frequent.
- Based on these frequent itemsets, association rules were generated.
- Metrics such as support, confidence, and lift were computed for evaluation purposes.
- Upon trying the model with a specific app name, the subsequent app recommendation was validated and found to be sensible according to user feedback.

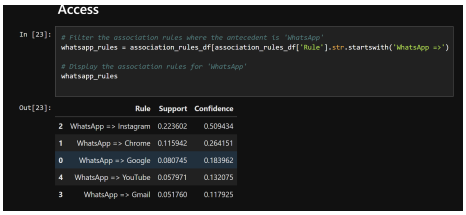

Figure 5: Next Possible Apps

4 Conclusion

Association rule mining, particularly in the context of mobile app usage, presents an intriguing avenue to understand and predict user behavior. By employing the SEMMA methodology in tandem with the Apriori algorithm, this research unveiled meaningful associations that can inform app recommendations. Such insights not only enhance user experience but can also drive engagement and retention. The validation of the model's recommendations, as corroborated by user feedback, further underscores its efficacy. Moving forward, it would be worthwhile to explore other algorithms and methodologies, refine the model with more extensive datasets, and possibly integrate temporal patterns to bolster the precision of recommendations.

5 References

1. Data Science PM. SEMMA. Available at: https://www.datascience-pm.com/semma/
2. Dataset: https://www.kaggle.com/datasets/atharvaarya25/phone-usage-dataset
3. Code: https://github.com/neel26desai/cmpe255$_{a}ssignment3/tree/main$