Frequent Itemsets and Association Rules

GitHub Link: https://github.com/Pritam0705/MACS-40123/tree/main/ITR1

Part 1: Analysis of Peer-Reviewed Papers

1. **A guided FP-Growth algorithm for mining multitude-targeted item sets and class association rules in imbalanced data**

   Author: Lior Shabtay, Philippe Fournier-Viger, Rami Yaari, Itai Dattner
   Published: Information Sciences, Volume 553, 2021
   Summary: This article presents a novel algorithm called Guided FP-Growth (G-FP-Growth) for mining multitude-targeted itemsets and class association rules in imbalanced datasets. The authors address the challenge of mining rare patterns in imbalanced data, a common issue in many real-world applications. The algorithm can mine multiple target classes simultaneously, allowing for more comprehensive pattern discovery. It introduces a novel tree structure called the Guided FP-tree, which enables efficient mining of rare patterns. The method can generate both frequent and rare class association rules, providing valuable insights into minority classes.

2. **Application of improved time series Apriori algorithm by frequent itemsets in association rule data mining based on temporal constraint**

   Author: Wang, C., Zheng, X.
   Published: Evolutionary Intelligence
   Summary: This article presents an improved time series Apriori algorithm based on temporal constraints for association rule data mining. The authors address the challenges of mining association rules in time series data, which is crucial for various applications such as financial analysis and weather forecasting. This enhancement allows for the discovery of more meaningful and temporally relevant patterns. The algorithm introduces a temporal constraint mechanism that considers the time dimension when generating frequent itemsets. This approach helps to identify patterns that are not only frequent but also temporally significant.

3. **Comparison Of Market Basket Analysis to Determine Consumer Purchasing Patterns Using Fp-Growth and Apriori Algorithm**

   Author: Ahmad Ari Aldino, Ahmad Ari Aldino, Setiawansyah, Sanriomi Sintaro, Ade Dwi Putra

Published: International Conference on Computer Science, Information Technology, and Electrical Engineering (ICOMITEE)

Summary: The primary goal of the research was to compare the performance and effectiveness of the FP-Growth and Apriori algorithms in analyzing consumer purchasing patterns through market basket analysis. The comparison revealed that the FP-Growth algorithm demonstrated superior performance compared to the Apriori algorithm. FP-Growth was found to be significantly faster in processing the dataset and generating results. The FP-Growth algorithm showed better efficiency in handling the given dataset. Both algorithms were able to identify frequent itemsets, but FP-Growth may have been more effective in discovering complex patterns. The findings of this study have important implications for businesses and researchers in the field of data mining and market analysis:

1. Algorithm Selection: FP-Growth may be the preferred choice for large datasets or time-sensitive applications due to its superior speed and efficiency.
2. Consumer Insights: The study demonstrates the value of these algorithms in uncovering valuable insights into consumer purchasing behaviors, which can inform marketing strategies and inventory management.

4. **Application of Association Rule Method Using Apriori Algorithm to Find Sales Patterns Case Study of Indomaret Tanjung Anom**

Author: Santoso, M. H.

Published: Brilliance: Research of Artificial Intelligence

Summary: This article presents a study on applying the Association Rule method using the Apriori algorithm to discover sales patterns at an Indomaret store in Tanjung Anom. The study revealed several important sales patterns and product associations within the Indomaret store: Frequently purchased item combinations were identified, providing insights into customer buying habits. Strong associations between certain products were discovered, indicating potential cross-selling opportunities. The analysis likely uncovered seasonal or time-based patterns in purchasing behavior.

Part 2: Interpretation

I have executed Apriori algorithm with the minimum threshold value [0.3, 0.5, 0.7, and 0.10].

For all thresholds, certain terms like “hindu,” “bangladesh,” and “medical” appear frequently as both antecedents and consequents. This indicates that these terms form a common pattern within the dataset, likely pointing to recurring misinformation themes involving Hindu, Bangladeshi, or RG Kar Medical college contexts.

Many rules have a confidence of 1.0, indicating that whenever the antecedent appears, the consequent always follows. High-confidence rules suggest strong co-occurrence, meaning that claims involving these antecedents reliably relate to the consequents, showing a direct connection in misinformation themes.

With context of two major events in August, the frequent itemsets and association rules become more insightful.

Example -

- (assaulted) → (hindu): This rule, with a confidence of 1.0 and a lift of 4.1, indicates that misinformation involving the word "assaulted" is almost always associated with the term "Hindu." This could reflect narratives linking the perceived vulnerability or targeting of Hindu minorities in Bangladesh following the political upheaval, possibly creating or amplifying fears among communities about safety and communal tensions.

- (bangladesh's) → (hindu): Similarly, this rule highlights how misinformation related to Bangladesh frequently pairs with themes surrounding the Hindu minority. This pattern likely reflects narratives painting Hindus in Bangladesh as at risk.

- (college) → (kar): This rule could relate to the RG Kar Medical College incident, where misinformation may have used terms like "college" and "Kar" frequently together to sensationalize or spread misleading details about the rape-murder case.

Major Themes -

- Itemsets like {hindu, bangladesh} or {hindus, bangladesh} show that a significant theme in the misinformation claims centers on communal tensions and the Hindu minority in Bangladesh.

- The itemset {rape, medical, murder, college, kar} connects terms that likely sensationalize the RG Kar Medical College case.

Summary

1. **What worked:**

- Testing different minimum support levels (e.g., 0.3, 0.5, 0.7) helped to find a balance between capturing meaningful frequent patterns and avoiding too many low-importance rules.
- Higher support levels (0.7) highlighted dominant themes without overwhelming us with minor associations.
- Knowledge of August's major events (Sheikh Hasina's resignation and the RG Kar Medical College incident) gave depth to the analysis by contextualizing the association rules and frequent itemsets. This context helped explain why certain terms frequently appeared together and how misinformation narratives were crafted around these events.

2. **What didn't work:**
    - Difficult to interpret results on large data
    - Prior knowledge regarding events is necessary to interpret associations and decide the themes of misinformation.

3. **Future revisions:**
    - Apply dynamic or variable thresholds based on prior knowledge or detected patterns. For instance, during peak misinformation events, setting a lower support would help to capture new narratives that may not have reached high frequencies.
    - Divide data into event-based or temporal segments (e.g., monthly or by news cycle) and apply separate FP-Growth analyses. This would highlight changes in misinformation patterns over time or during specific events, yielding more targeted insights.
    - Use topic modeling alongside FP-Growth to identify underlying topics or narratives within frequent itemsets. This would help capture the nuance that association rules alone might miss, especially for interpreting multi-word phrases and complex themes.

Part 3: Social, Cultural, and Behavioral Implications of Findings

The patterns uncovered in the analysis of misinformation in India reveal deep-seated social, cultural, and behavioral dynamics that go beyond the empirical results. By showing how misinformation spreads in the context of particular themes, such as religion, politics, and identity, the findings indicate that misinformation is not only a consequence of technological networks but also a reflection of society's social and cultural fabric.

Social Implications

The recurring misinformation themes surrounding Hindu minorities, regional identities, and political events expose how misinformation exploits sensitive social divisions. Misinformation, by targeting and reinforcing these themes, amplifies social polarization, fostering mistrust among communities and increasing the potential for conflict. This use of misinformation reflects a broader social phenomenon where information is used as a tool for influence, steering public sentiment by appealing to specific community loyalties.

Behavioral Implications

On a behavioral level, the findings reveal that misinformation often spreads through patterns of confirmation bias and social identity reinforcement. When misinformation aligns with an individual's social identity or pre-existing beliefs, they are more likely to share it, contributing to its viral potential.

These findings align with theories on misinformation and social identity, such as *Social Identity Theory*, which posits that people are inclined to accept information that supports their in-group. Additionally, *Cultivation Theory*, which suggests that repeated exposure to certain narratives shapes public perception, is evidenced by the prominence of repeated themes like communal tensions.

Conclusion

The analysis reveals that misinformation in India capitalizes on social divisions, cultural narratives, and identity biases, deepening polarization and reshaping trust in information sources. By aligning with existing beliefs and cultural scripts, misinformation spreads more easily, reinforcing echo chambers and societal divides. Addressing misinformation thus requires strategies that consider not only the content but also the social and cultural dynamics that fuel its spread.