Research Proposal: Addressing the Semantic Gaps in Association Rule Mining via LLM-Augmented Hypothesis Generation and Validation

Group 11

KDD Course Project

Instructor: Raymond Wong

September 25, 2025

1. Background and Motivation

Traditional Association Rule Mining (ARM) has been widely adopted to uncover co-occurrence patterns in data. However, it faces several fundamental limitations. First, ARM is purely frequency-driven, which leads to a lack of semantic depth: discovered rules often reflect surface-level co-occurrence without understanding synonymy, hierarchy, or implicit causal relationships. Second, ARM struggles with data sparsity and long-tail phenomena, thereby ignoring rare but scientifically valuable associations, such as infrequent gene–drug interactions. Third, ARM is constrained within a single dataset or domain, making it difficult to uncover cross-domain knowledge transfer or hidden analogies across disciplines. Finally, although the extracted rules are syntactically interpretable, they lack explanatory power, offering no insight into why a particular association holds.

Large Language Models (LLMs) present an opportunity to address these shortcomings. By leveraging pretrained semantic representations, LLMs can enrich rule mining with deeper semantic understanding, allowing surface itemsets to be mapped onto conceptually meaningful entities. Moreover, LLMs can hypothesize low-frequency but plausible associations, mitigating the long-tail limitation. Their broad pretraining further enables discovery of cross-domain analogies, bridging knowledge silos between different scientific areas. Additionally, LLMs can generate natural language rationales that enhance interpretability by providing potential causal or mechanistic explanations for each rule. This motivates the integration of LLMs with traditional ARM to build a hybrid framework that combines statistical rigor with semantic richness, advancing the reliability and usefulness of association discovery.

2. Research Objectives

The primary objectives of this research are as follows:

- Propose a systematic framework that integrates Large Language Model (LLM)-generated semantic knowledge with traditional association rule mining, addressing key limitations such as semantic gaps, long-tail neglect and limited interpretability.
- Assess the added value of LLM-augmented associations in terms of semantic richness, explanatory depth and cross-domain applicability, relative to conventional frequency-based approaches.
- Develop and employ novel evaluation metrics including semantic coherence, contextual relevance and interpretability that extend beyond classical measures such as support, confidence and lift.
- Conduct empirical studies on both real-world and synthetic datasets to systematically evaluate the effectiveness, robustness and generalizability of the proposed approach.

3. Related Work Review

3.1. Classical Association Rule Mining and Limitations

Association Rule Mining (ARM), with foundational algorithms such as Apriori [1] and FP-Growth [2], has been widely adopted for discovering frequent co-occurrence patterns. However, these approaches are frequency-driven and thus exhibit critical shortcomings: lack of semantic depth, inability to capture rare but meaningful associations, rule explosion and limited explanatory power. Extensions to numerical association rule mining attempted to handle continuous attributes more effectively, but the semantic gap remains.

3.2. Semantic and Knowledge-Augmented Data Mining

To overcome the semantic limitations of ARM, researchers explored incorporating external knowledge, ontologies or semantic embeddings into the mining process. Surveys such as [3] highlight ontology-guided methods that integrate synonymy, hypernymy and domain-specific hierarchies into association discovery. In biomedical domains, ontology-enriched mining has been used for gene–disease and drug–symptom associations, showing that semantic priors improve interpretability and relevance. More recently, LLM-based approaches have been applied for ontology enrichment, automatically extracting is-a relationships and semantic triples from unstructured texts, demonstrating the potential of LLMs in knowledge augmentation.

3.3. LLMs and Rule/Knowledge Generation

Recent work has investigated the synergy of symbolic reasoning and LLMs. *ChatRule* [4] introduces an LLM-based framework for logical rule generation in knowledge graphs, combining prompt-driven rule generation with data-driven filtering. Other works, such as [5], distill interpretable first-order logic rules and inject them into LLM prompts, enhancing reasoning. Similarly, LLMs have been applied to property graph rule mining, where graph structures are encoded as text and candidate rules are inferred via prompting. These efforts highlight the promise of hybrid symbolic–neural pipelines but often lack rigorous statistical validation against spurious or "hallucinated" rules.

3.4. LLMs as Semantic Embedding Models

Beyond rule generation, LLMs have demonstrated effectiveness as embedding models for semantic representation. Recent surveys show that LLM embeddings outperform traditional word embeddings in semantic clustering and retrieval tasks [6]. This suggests that embedding-based association discovery can benefit from LLM-derived semantic spaces, aligning frequent itemsets with richer conceptual meanings.

3.5. Spatiotemporal and Semantic Association Mining with LLMs

The recently published work *Advancing Large Language Models for Spatiotemporal and Semantic Association Mining of Similar Environmental Events* [7] proposes a two-stage retrieval and re-ranking framework that integrates LLM embeddings with spatiotemporal signals for environmental event recommendation. Their Geo-Time Re-ranking (GT-R) model combines semantic similarity, category relevance, spatial proximity and temporal association, achieving superior performance over traditional dense retrieval baselines. This study underscores the importance of extending LLM-based semantic association beyond pure text to incorporate spatial and temporal dimensions, a principle that resonates with the goal of enhancing association mining through multi-faceted semantics.

3.6. Research Gap and Positioning

Most existing LLM–rule mining efforts focus on knowledge graphs or retrieval tasks, while semantic augmentation of classical ARM on transactional or text-based datasets remains underexplored. Prior work rarely integrates LLM-generated candidate associations with statistical significance testing, leaving the hallucination problem unsolved. Moreover, evaluation metrics are often confined to precision and recall, neglecting semantic coherence and interpretability. Our proposed research aims to address these gaps by combining:

1. LLM-driven candidate association generation and semantic expansion;
2. Statistical validation (e.g., chi-square, PMI) to filter hallucinations;

3. Novel evaluation metrics emphasizing semantic coherence and explanatory depth;
4. Application to scientific literature and cross-domain discovery, moving beyond the transactional ARM paradigm.

4. Methodology / Research Design

4.1. Data Sources

We will select one or more datasets where semantic relationships are important (e.g., product co-purchase data, medical condition-symptom pairs, research abstracts).

4.2. Baseline

We will implement classic association rule mining algorithms as baselines.

4.3. LLM Integration

We propose two modes of enhancement:

- **Post-hoc enrichment:** Re-ranking mined rules with LLM-based semantic similarity.
- **LLM-guided candidate generation:** Using LLMs to propose or validate item pairs before rule mining.

4.4. Evaluation Strategy

Evaluation metrics will include:

- Quantitative: support, confidence, lift, semantic similarity scores
- Qualitative: human evaluation of relevance and interpretability

5. Expected Results and Contribution

We expect that LLM-enhanced association rules will:

- Improve semantic relevance and coverage of discovered patterns
- Provide better interpretability for end users
- Enable cross-domain generalization by leveraging pretrained knowledge

Our contribution will be a prototype system and experimental study demonstrating these enhancements.

6. Project Timeline and Members Information

Week	Milestone / Task
Week 1	Finalize topic, literature review assignment to team members
Week 2	Prepare baseline implementations (Apriori, FP-Growth)
Week 3	LLM prompt design and API integration
Week 4	Dataset preparation and rule mining experiments
Week 5	Evaluation (quantitative + qualitative), ablation tests
Week 6	Report writing, slides preparation, final presentation

Table 1: Project timeline

Name	Student ID	Research / FYP Topic
Man Chun Hei William	20708177	Speaking Chatbot Simulation
Cheng Ho Yin	20958239	Fly Drones
Guoying Lu	21270828	Artificial Intelligence Governance
Wong Cheuk Yuen	20860468	Constrained Shortest Path Queries
Sirui Han	12281981	Sovereign AI Model
Tsang Wai Lok	20866979	AI-Centric Networking

Table 2: Project Team Members and Their Research Topics

7. Declaration

We hereby declare that this project has been carried out solely as part of the requirements of the COMP 5331 course.

All research activities, implementations, and deliverables associated with this work are restricted to the academic scope of the course, and are not intended for use in any external research, publication, or commercial application. The project is conducted exclusively within the context of the course and does not overlap with other academic, professional, or personal projects.

References

[1] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," *Proceedings of the 20th International Conference on Very Large Data Bases (VLDB)*, pp. 487–499, 1994.

[2] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation: A frequent-pattern tree approach," *ACM SIGMOD Record*, vol. 29, no. 2, pp. 1–12, 2000.

[3] W. Li, C.-R. Shyu, et al., "Semantic data mining: A survey of ontology-based approaches," *Knowledge and Information Systems*, vol. 47, no. 3, pp. 1–42, 2015.

[4] K. Sun, B. Xu, H. Zhang, et al., "Chatrule: Mining logical rules with large language models for knowledge graph reasoning," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

[5] W. Zhang, L. Chen, and S. Wang, "Learned-rule-augmented generation for large language models," *Transactions of the Association for Computational Linguistics*, 2024.

[6] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *Proceedings of EMNLP*, 2019.

[7] Y. Tian et al., "Advancing large language models for spatiotemporal and semantic association mining of similar environmental events," *International Journal of Geographical Information Science*, 2025, Preprint available.