Frequent Itemsets and Association Rules for JobSkills Dataset

Salima Tankibayeva

November 15, 2024

1 Introduction

In this report, we applied the PCY (Park-Chen-Yu) algorithm to discover frequent itemsets from a job skills dataset. We then used these itemsets to generate association rules based on minimum support and confidence thresholds. The results were visualized using various techniques, including heatmaps and network graphs, to explore patterns and relationships between different skills.

2 Methodology

2.1 Data Preprocessing

We started with a dataset containing job skills information. The dataset was preprocessed as follows:

- Missing values in the "job_skills" column were filled with an empty string.
- Each job skill entry was split by commas to create a list of skills.
- A normalization step was performed to standardize skill names.

2.2 PCY Algorithm

The PCY algorithm was used to find frequent itemsets:

- **Step 1:** Convert the transactions into a sparse matrix for efficient computation.
- **Step 2:** Count single items and store frequent ones.
- **Step 3:** Generate candidate itemsets by examining pairs of items and using a hash function to reduce candidate space.
- **Step 4:** Generate larger itemsets by combining smaller frequent itemsets.

Frequent itemsets were identified based on the minimum support threshold. The frequent itemsets were used to generate association rules with a minimum confidence threshold.

2.3 Implementation

Normalization Challenges: The normalization of skill titles (e.g., mapping "communication skills" and "communication" to a common label) mitigated redundancy in the analysis. However, such preprocessing requires careful attention to avoid misrepresentation or loss of nuance.

Scalability and Efficiency: The implementation of the PCY algorithm allowed efficient analysis of a large dataset, demonstrating its scalability for practical applications. The use of hash buckets and bitmaps optimized computational resources during the frequent itemset generation.

2.4 Association Rule Generation

Association rules were generated from the frequent itemsets using the following criteria:

- A rule was generated for itemsets with more than one item.
- Rules were only retained if the confidence met the minimum confidence threshold.

2.5 Visualization

The results were visualized using the following techniques:

- **Heatmap:** A heatmap was plotted to visualize the confidence values of the generated association rules.
- **Network Graph:** A network graph was created to show the relationships between different job skills based on the generated rules.

3 Results

3.1 Frequent Itemsets

The frequent itemsets and their support values are displayed below:

Itemset	Support
{attention to detail}	133,975
{collaboration}	87,116
{communication}	566,092
{customer service}	278,102
{data analysis}	81,964
{high school diploma}	67,267
{interpersonal skills}	100,267
{inventory management}	71,911
{leadership}	185,187
{microsoft office suite}	75,531
{nursing}	88,015
{organizational skills}	75,274
{patient care}	99,926
{problem solving}	278,361
{project management}	121,563
{sales}	93,031
{teamwork}	227,609
{time management}	142,911
{training}	83,656

Table 1: Frequent Itemsets with Support

The analysis highlights the importance of core transferable skills like *communication*, *problem solving*, and *teamwork*, which are common across many professions.

Significant skill combinations include:

- {Communication, Problem Solving} with a support of 219,241.
- {Communication, Customer Service} with a support of 189,532.
- {Communication, Teamwork} with a support of 181,643.

These combinations are frequent in job postings, reflecting their practical relevance in the workplace.

3.2 Association Rules

The top association rules generated from the frequent itemsets include the following:

- {interpersonal skills} → {communication}, confidence: 0.75
- {problem solving} → {communication}, confidence: 0.79
- {communication} → {problem solving}, confidence: 0.39
- {collaboration} → {communication}, confidence: 0.75
- {teamwork} → {customer service}, confidence: 0.40
- {customer service} → {teamwork}, confidence: 0.33

3.3 Bidirectional Relationships

Communication & Teamwork

- Communication → Teamwork (confidence: 0.32)
- Teamwork → Communication (confidence: 0.80)

This suggests that teamwork heavily relies on communication, but not all communicators necessarily work in teams.

Problem Solving & Leadership

- Leadership → Problem Solving (confidence: 0.48)
- Problem Solving → Leadership (confidence: 0.32)

Leadership skills often encompass problem-solving, though not all problem-solvers take on leadership roles.

3.4 Bidirectional Relationships

Communication & Teamwork

- Communication → Teamwork (confidence: 0.32)
- Teamwork → Communication (confidence: 0.80)

This suggests that teamwork heavily relies on communication, but not all communicators necessarily work in teams.

Problem Solving & Leadership

- Leadership → Problem Solving (confidence: 0.48)
- Problem Solving → Leadership (confidence: 0.32)

Leadership skills often encompass problem-solving, though not all problem-solvers take on leadership roles.

3.5 Cross-Functional Skills

Customer Service → Communication (confidence: 0.68) Effective customer service is closely tied to communication skills.

Sales → Communication (confidence: 0.78) Communication is essential in sales roles, as expected.

Project Management → Communication (confidence: 0.62) Projects require constant communication for successful execution.

3.6 Three-Skill Associations

- Customer Service, Teamwork → Communication (confidence: 0.87): Roles combining customer service and teamwork almost always necessitate strong communication.
- Problem Solving, Customer Service → Communication (confidence: 0.90): Problem-solving and customer-facing roles are closely tied to communication.

3.7 Network Graph of Association Rules

The network graph shows the relationships between job skills based on association rules. Each node represents a skill, and each edge between nodes represents a rule, with the edge weight corresponding to the confidence of the rule.

3.8 Network Graph of Association Rules

The network graph shows the relationships between job skills based on association rules. Each node represents a skill, and each edge between nodes represents a rule, with the edge weight corresponding to the confidence of the rule.

4 Discussion

The findings provide valuable insights into job market trends and skill requirements:

Importance of Communication Skills: Communication emerged as the most frequent and interconnected skill, emphasizing its universal demand across industries. Its high-confidence relationships with *teamwork*, *customer service*, and *problem solving* reflect its role as a foundational skill.

Interdisciplinary Skill Sets: The frequent co-occurrence of skills like *communication*, *problem solving*, and *teamwork* highlights the growing emphasis on interdisciplinary capabilities. For instance, association rules sug-

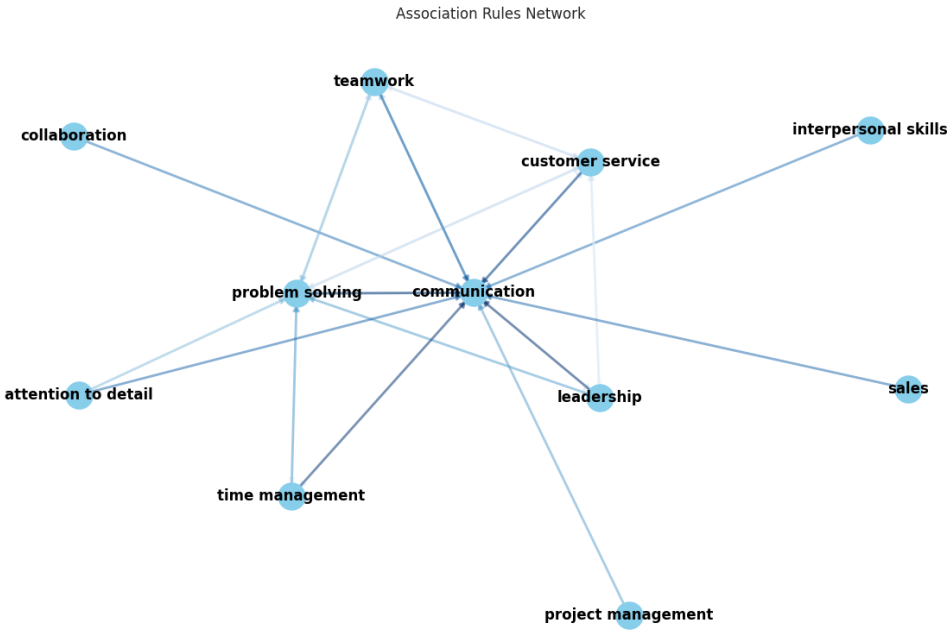

Figure 1: Network Graph of Association Rules

gest that a combination of soft skills like teamwork and problem solving significantly correlates with communication abilities.

Heatmap of Association Rules The heatmap visualizes the confidence values of the generated association rules. Each rule is represented as a cell, with higher confidence values in darker colors.

4.1 Limitations and Future Work

Limited Contextual Understanding: The analysis does not consider the contextual nuances of skills within specific job roles.

Sequence Analysis: The sequential dependencies between skills (e.g., transitions in learning or career progression) could be explored further using Markov models or transition matrices.

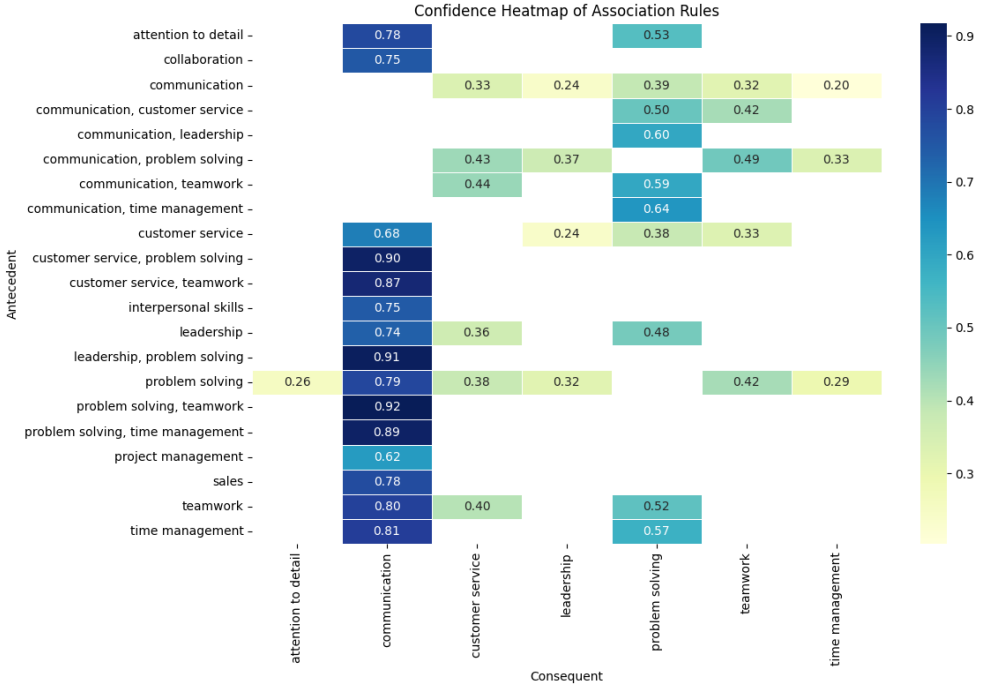

Figure 2: Confidence Heatmap of Association Rules

Dynamic Trends: Incorporating temporal data could provide insights into evolving skill demands over time.

5 Conclusion

This study demonstrates the utility of data mining techniques like the PCY algorithm in extracting meaningful patterns from job skills data. The frequent itemsets and association rules highlight the prominence of key skills and their interrelationships, offering actionable insights for employers, educators, and policymakers to align workforce development with market demands. Future work can expand on these findings by incorporating more contextual and temporal data to deepen the analysis.

6 Declaration

"I/We declare that this material, which I/We now submit for assessment, is entirely my/our own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of my/our work, and including any code produced using generative AI systems. I/We understand that plagiarism, collusion, and copying are grave and serious offences in the university and accept the penalties that would be imposed should I engage in plagiarism, collusion or copying. This assignment, or any part of it, has not been previously submitted by me/us or any other person for assessment on this or any other course of study."