Market-basket Analysis

Oumaima El Menni

June 2024

1 Introduction

In today's competitive job market, understanding the skills in demand is crucial for job seekers, employers, and educational institutions. The LinkedIn Jobs & Skills dataset provides a rich source of information about the skills required across various job postings, offering valuable insights into current market trends and skill requirements.

This project aims to leverage market-basket analysis, specifically the Frequent Pattern Growth (FPGrowth) algorithm, to uncover frequent itemsets and association rules within the job skills listed in the LinkedIn dataset. By treating job postings as transactions and the listed skills as items, we can identify which skills commonly appear together and how often they do so.

The primary objective of this analysis is to find frequent combinations of skills and to understand the relationships between them. These insights can help job seekers tailor their skill development to meet market demands, enable employers to craft more effective job descriptions, and assist educational institutions in designing relevant curricula.

Given the size and complexity of the dataset, we utilize Apache Spark for efficient data processing and analysis. Spark's distributed computing capabilities allow us to handle large datasets and perform complex computations in a scalable manner.

This report will detail the methodology used to preprocess the data, apply the FPGrowth algorithm, and interpret the resulting frequent itemsets and association rules. The findings will highlight significant skill combinations and their implications, providing actionable insights for various stakeholders in the job market.

2 Dataset Description

2.1 Source

The dataset used for this analysis is the "1-3M LinkedIn Jobs and Skills 2024" dataset, which was obtained from Kaggle. This dataset contains job postings from LinkedIn, including various attributes related to job details and required skills.

2.2 Content

The dataset consists of two main columns:

- job link: A unique identifier for each job posting, represented as a URL link to the job listing.
- job skills: A string containing a list of skills required for the job, with skills separated by commas.

2.3 Size

The dataset contains 1,294,374 job postings, making it approximately 2GB in size. This substantial size necessitates the use of a distributed computing framework like Apache Spark to efficiently handle and process the data.

3 Data Preprocessing

To ensure that it is in the appropriate format for analysis, the data preprocessing stage involves handling missing values, transforming the "job skills" column into an array of individual skills (this transformation allows us to treat each job posting as a collection of skills, essential for subsequent analysis), and performing basic data quality checks:

- Null values and empty strings in the "job skills" column were removed to ensure data integrity.
- Text was normalized by converting it to lowercase, removing leading and trailing white space, and replacing multiple spaces with a single space.
- Punctuation marks are removed from the "job skills" column.

4 Methodology

The methodology employed in this analysis involves utilizing market-basket analysis techniques to extract meaningful insights from the LinkedIn Jobs & Skills" dataset. Market-basket analysis is a data mining technique commonly used in retail and e-commerce to identify associations between items purchased together. In the context of job postings and required skills, market-basket analysis helps uncover frequent combinations of skills and understand the relationships between them.

4.1 FP-Growth Algorithm

The FP-Growth (Frequent Pattern Growth) algorithm is a popular technique for mining frequent itemsets from transactional datasets. It involves the following key components:

4.1.1 FP-tree Construction

The FP-tree (Frequent Pattern Tree) is a compact data structure that represents the frequency of items in a transactional dataset. It consists of nodes and links, where each node represents an item and its frequency.

4.1.2 Support Count

The support count of an itemset X is defined as the number of transactions in the dataset that contain X. Mathematically:

$$\text{Support}(X) = \frac{|\{T_i \mid X \subseteq T_i, T_i \in D\}|}{|D|}$$

where D is the transaction dataset, and T_i is an individual transaction.

4.1.3 Conditional Pattern Base

For each frequent item, a conditional pattern base is constructed, which is a sub-dataset containing the prefix paths that co-occur with the item.

4.1.4 Recursive Mining

The FPGrowth algorithm recursively mines the FP-tree by constructing conditional FP-trees for the conditional pattern bases and extracting frequent itemsets. This process continues until no more frequent itemsets can be found.

4.2 Relevance for the Task

The decision to employ the FP-Growth algorithm in this project stems from its aptness for analyzing the LinkedIn Jobs & Skills dataset, which comprises over a million job postings. Given the scale of the dataset, FP-Growth's efficiency in handling large transactional data without the need to generate candidate itemsets explicitly is particularly advantageous. This algorithm excels in compressing sparse data into a compact FP-tree representation, facilitating the discovery of frequent itemsets and association rules even amidst the diverse and numerous skill combinations present in job postings. Moreover, FP-Growth's scalability ensures that it can process the dataset within a reasonable timeframe, enabling timely insights into skill patterns and their co-occurrences across various job postings. By leveraging FP-Growth, the project aims to uncover meaningful insights into prevalent skill requirements and associations in job descriptions, thereby providing valuable information for job seekers, employers, and educational institutions navigating the dynamic job market landscape.

Parameters: The choice of `minSupport` and `minConfidence` values is critical for the effectiveness of the FPGrowth algorithm.

- `minSupport`: This parameter determines the minimum support threshold for an itemset to be considered frequent. It represents the minimum proportion of transactions in which an itemset must appear to be deemed frequent. A lower `minSupport` value results in more itemsets being identified as frequent, potentially leading to longer computation times and more memory usage.
- `minConfidence`: This parameter specifies the minimum confidence threshold for association rules to be generated from the frequent itemsets. Confidence measures the strength of an association rule and represents the probability of the consequent item appearing in a transaction given that the antecedent item(s) are present. A higher `minConfidence` value filters out weaker association rules, ensuring that only significant rules are considered.

For this project, we chose a minSupport value of 0.05 because we are interested in identifying itemsets that appear in at least 5% of the total transactions. This value strikes a balance between capturing frequent patterns while avoiding overly common itemsets that may not provide meaningful insights. Similarly, a minConfidence value of 0.1 was selected because we are interested in association rules that have a confidence level of at least 10%. This threshold ensures that the discovered rules are statistically significant and have a strong likelihood of being valid.

5 Results

5.1 Frequent Itemsets

The analysis of the LinkedIn Jobs & Skills dataset using the FPGrowth algorithm revealed several frequently occurring skills and combinations of skills in job postings. The top 20 most frequent itemsets are presented in Table 1.

Items	**Frequency**
Communication	366,270
Customer Service	276,788
Teamwork	226,212
Communication Skills	195,249
Leadership	183,839
Problem Solving	146,164
Time Management	142,148
Teamwork, Communication	139,320
Customer Service, Communication	139,202
Attention to Detail	132,784
Problemsolving	127,174
Project Management	120,177
Leadership, Communication	117,281
Interpersonal Skills	99,943
Patient Care	99,454
Problem Solving, Communication	94,238
Sales	92,744
Teamwork, Customer Service	91,091
Problemsolving, Communication	89,846
Nursing	87,833

Table 1: Top 20 Most Frequent Itemsets in Job Postings

These top 20 skills give us a clear view of what employers are looking for in job candidates. Communication comes first, showing how important it is for any job. Customer service and teamwork follow closely, telling us that working well with others is key. It's interesting to see 'communication skills' mentioned separately, showing they're a big deal on their own. And leadership and problem-solving skills are also high up, showing they're really valued. The list also includes specific skills like project management and patient care, showing the variety of things employers need. This helps both job seekers and employers understand what's important in today's job market.

Figure 1: WordCloud of the most frequent skills

5.2 Association Rules

Using the frequent itemsets, we derived several association rules that highlight interesting relationships between different skills. Table 2 shows the top 20 association rules based on their confidence levels.

5.2.1 Metrics Interpretation

- **Antecedent**: The initial item or itemset in an association rule. It is the "if" part of the rule. For example, in the rule $[A] \rightarrow [B]$, A is the antecedent.
- **Consequent**: The item or itemset that is likely to appear given the antecedent. It is the "then" part of the rule. For example, in the rule $[A] \rightarrow [B]$, B is the consequent.
- **Confidence**: A measure of the reliability of the rule. It is the proportion of transactions containing the antecedent that also contain the consequent. High confidence means the rule is often true.
- **Lift**: A measure of how much more likely the consequent is to appear when the antecedent is present compared to when it is not. A lift greater than 1 indicates a positive correlation between the antecedent and consequent.
- **Support**: The proportion of transactions in the dataset that contain both the antecedent and the consequent. It reflects how frequently the rule is applicable in the dataset.

5.2.2 Comment

The analysis of the top 20 association rules reveals intriguing patterns in the LinkedIn Jobs & Skills dataset. These findings shed light on the intricate interplay between different job skills within professional settings. Notably, communication emerges as a pivotal skill across various domains, as evidenced by the high confidence levels in associations like "problemsolving" and "communication" or "leadership" and "communication". This underscores the vital role effective communication

Antecedent	Consequent	Confidence	Lift	Support
problemsolving	communication	0.706	2.497	0.069
problem solving	communication	0.645	2.278	0.073
leadership	communication	0.638	2.254	0.091
teamwork	communication	0.616	2.176	0.108
time management	communication	0.593	2.097	0.065
customer service	communication	0.503	1.777	0.108
attention to detail	communication	0.495	1.751	0.051
teamwork	customer service	0.403	1.883	0.070
communication	teamwork	0.380	2.176	0.108
communication	customer service	0.380	1.777	0.108
leadership	customer service	0.362	1.694	0.051
customer service	teamwork	0.329	1.883	0.070
communication	leadership	0.320	2.254	0.091
communication	problem solving	0.257	2.278	0.073
communication	problemsolving	0.245	2.497	0.069
customer service	leadership	0.241	1.694	0.051
communication	time management	0.230	2.097	0.065
communication	attention to detail	0.180	1.751	0.051

Table 2: Top 20 Association Rules by Confidence

plays in problem-solving and leadership contexts. Moreover, the frequent coupling of teamwork with communication skills suggests the significance of clear communication channels in collaborative efforts. The strong confidence observed in rules linking customer service with communication and teamwork underscores the importance of interpersonal skills in client-facing roles. These insights underline communication as a cornerstone skill, crucial for navigating diverse professional scenarios and enhancing overall job performance.

6 Conclusion

In conclusion, our analysis of the LinkedIn Jobs & Skills dataset yielded several noteworthy findings. Firstly, we observed that communication, customer service, teamwork, and leadership are prevalent skills demanded across various job postings, indicating their significance in today's workforce. Furthermore, association rules uncovered insightful correlations, such as the strong association between problem-solving and leadership, and the frequent pairing of communication with customer service. These results not only provide valuable guidance for job seekers in tailoring their applications but also offer crucial insights for employers in crafting compelling job descriptions. By leveraging these findings, stakeholders can enhance their recruitment strategies, foster better talent matches, and ultimately drive success in their respective endeavors within the job market.

Declaration

I declare that this material, which I now submit for assessment, is entirely my 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 work, and including any code produced using generative AI systems. I 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 or any other person for assessment on this or any other course of study.