Algorithms for Massive Data Project
Market Basket Analysis

Mahnaz Taheri
matriculation number: 989363

July 2024

1 Introduction

In this project, Market Basket Analysis has been applied on the dataset called Linkedin Jobs & Skills (2024), considering the skills as items and the sets of skills defined for each row as baskets.

To implement the Apriori algorithm with our data, PySpark is chosen for its ability to process tasks in parallel. PySpark, a Python library for Apache Spark, spreads out computations across multiple machines, which is key for handling large datasets. This approach helps uncover relationships and patterns among different sets of skills, similar to how market basket analysis identifies associations in transactional data.

2 Dataset and Pre-processing

The dataset, Linkedin Jobs & Skills (2024), is obtained from kaggle. This dataset contains almost 1.3 million job listings scraped from LinkedIn in 2024 which is a popular professional networking site with millions of job postings.

There are two columns in our dataset which are: job_link and job_skills. Where the job_link is being ignored during the analysis since it does not serve in project's purpose. Hence, the job_skills column is the only column considered for the analysis where in each row, exist a set of strings taken into account as baskets. While each string is considered as a skill or an item.

The dataset is made of 1,287,105 rows. For simplifying the analysis, a sample has been obtained with 12,868 rows or baskets where the total number of items or skills is 457,130 (the distinct number of skills is 19,230).

In the preprocessing stage, several transformations are applied to the data. Each skill, which consists of strings, was first tokenized to break it down into individual words. Stop words, such as common words like 'the' and 'is', were removed. Next, lemmatization was applied to reduce words to their base form. This preprocessing helps ensure that we analyze skills effectively by focusing on the essential content.

At the end, each skill, represented as a string, is converted to a unique number to enhance the algorithm's functionality.

3 Algorithm

Market Basket Analysis (MBA) is a data mining technique used to uncover relationships between items in large datasets, commonly applied in retail to understand customer purchasing patterns. By analyzing transaction data, MBA identifies associations or "rules" between products. In our dataset, this method is employed to identify frequent itemsets and generate association rules.

Association rules are patterns found in data that reveal relationships between items indicating that certain items or events tend to occur together. In the context of a LinkedIn dataset, association rules reveal meaningful connections between different skills.

Below the steps for building the Association Rules are defined:

- Support Calculation: This measures the frequency with which an item (or a combination of items) appears in the dataset. We calculate the support for each itemset by counting the number of transactions containing that itemset and dividing it by the total number of transactions.

- Generate Candidate Itemsets: The candidates are potential itemsets that might be frequent. Start with single items and generate candidate itemsets of higher sizes based on the frequent itemsets found in the previous step.

- Confidence: It measures the reliability of the association rule. It is the ratio of the support of the combined itemset to the support of the antecedent itemset.
  - Antecedent (X): The initial itemset or condition in a rule.
  - Consequent (Y): The itemset or outcome that follows from the antecedent in a rule.

  The confidence of a rule $X \rightarrow Y$ measures how often items in Y appear in transactions that contain X:

  $$Confidence(X \rightarrow Y) = \frac{Support(X \cup Y)}{Support(X)}$$

  - Support($X \cup Y$): The proportion of transactions that contain both X and Y.
  - Support(X): The proportion of transactions that contain X.

- Lift: Measures how much more often skills occur together than we would expect if their occurrence were independent.

  The lift of a rule $X \rightarrow Y$ is defined as:

  $$Lift(X \rightarrow Y) = \frac{Support(X \cup Y)}{Support(X) \times Support(Y)}$$

  - Support($X \cup Y$): The proportion of transactions that contain both X and Y.
  - Support(X): The proportion of transactions that contain X.
  - Support(Y): The proportion of transactions that contain Y.

  If the lift of an itemset is greater than 1, it indicates that the items of the itemset are more likely to appear together.

  If it is equal to 1, then the items are independent of each other. This means the presence of one item has no effect on the presence of another.

  Finally, if it is less than 1, it means that the items are less likely to appear together. This often indicates a negative association where the presence of one item discourages the presence of another.

4 Implementation

To implement Apriori algorithm, the following steps have been taken:

- Compute Frequencies: Calculate the frequency of each skill, which is the number of occurrences of each item. Only consider skills appearing in at least 10% of all baskets will be considered frequent.

- Normalize Frequencies: Normalize these frequencies by dividing each by the total number of baskets.

- Generate Candidates: Determine the combinations of skills based on the desired itemset size. Start with pairs (two-item combinations), finding all pairs (doubletons) that occur in the baskets and compute the support for each.

- Extend to Larger Itemsets: Repeat the process for larger itemsets, such as triples (three-item combinations), finding all triples that occur in the baskets and computing their support.

Now that all frequent itemsets are identified, the next step is to generate association rules. The confidence and lift for each rule is calculated and then are filtered based on minimum confidence threshold (70%).

5 Results

To clearly present the results of the algorithm applied to the LinkedIn skills dataset, Table 1 has been constructed including antecedent, consequent, support, confidence and lift.

Table 1: Association Rules for Linkedin Dataset

Antecedent	Consequent	Support	Confidence	Lift
time	management	0.1424	0.8336	1.6614
skill service	customer	0.1049	0.7675	2.7424
detail	attention	0.1052	0.9191	8.6637
communication problem	solving	0.1063	0.9253	7.7592
time	communication management	0.1207	0.7067	2.0587

The support of each itemset indicates the percentage of occurrences of that itemset within the entire dataset. For example, the itemset time, management has a support of 0.1424, meaning that this pair appears in 14% of all transactions in the dataset. The confidence for this pair is 83%, indicating that 83% of transactions containing ”time” also contain ”management.” The lift for this pair is 1.66, which is greater than 1 suggesting a positive association meaning that the items co-occur more frequently than expected by chance.

“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.“