Market Basket Analysis with Apriori Algorithm

UNIVERSITÀ
DEGLI STUDI
DI MILANO

Course: Privacy, data protection, and massive data analysis in emerging scenarios

Module: Algorithms for massive datasets

Name: Ahmet Emre Iskender

Student ID: 11001A

INDEX

1)Introduction

Market Basket Analysis is a widely used data mining technique to construct associations between the items that are generally purchased together. Market Basket Analysis is a technique that market retailers and even the e-commerce industry participants are using to grow their revenues. It is very important to use to understand the consumer buying patterns so that by using Market Basket Analysis approach they bring the items together and create bundles from the items that are usually bought by the customer together.

Considering the fact that for a retailer following up the consumer data is highly valued and generally these data are massive. Thus, an algorithm is needed to conduct Market Basket Analysis. There are various algorithms to conduct Market Basket Analysis such as Apriori or PCY but since in this experiment the used algorithm is Apriori, we will be focusing on the details of the Apriori algorithm in the following sections. As indicated above, Apriori algorithm is used to conduct Market Basket Analysis which also has parameters such as minimum support to generate frequent itemsets and confidence threshold to create association rules from the frequent items. In order to apply Apriori algorithm, the dataset structure should be either lists or dictionaries considering the fact that it focuses on the transaction bundles to understand which items are more associated together.

2)Data Exploration & Data Preprocessing

The data that is used in this report is about 1.3 million LinkedIn job postings and the skills related to these jobs. Thus, there are only 2 columns and 1.3 million rows. In the first column called “job_link” there are only the links of the jobs and in the second column named “job_skills” there are skills for each job postings separated by commas. Since the analysis will be focused on the “job_skills”, as the first thing the null values must be checked and if there is any they must be removed. While checking the null values, it is seen that there were 20135 null values under the job_skills column so they have been removed from the dataset.

Secondly, since the analysis was made upon Python 3, the rows of job_skills column are constructed into sets. Then, the job_skill column was separated into a new data frame and after each set was stored in a single list. Furthermore, in order to better understand the data, the maximum and minimum set sizes are found as 276 and 1 meaning that there is a job posting which requires 276 different skills for the applicants to be admitted for it. In order to drill down into details and better understand the overall picture of the set sizes the distribution of set sizes has been plotted and can be seen from Figure 1.

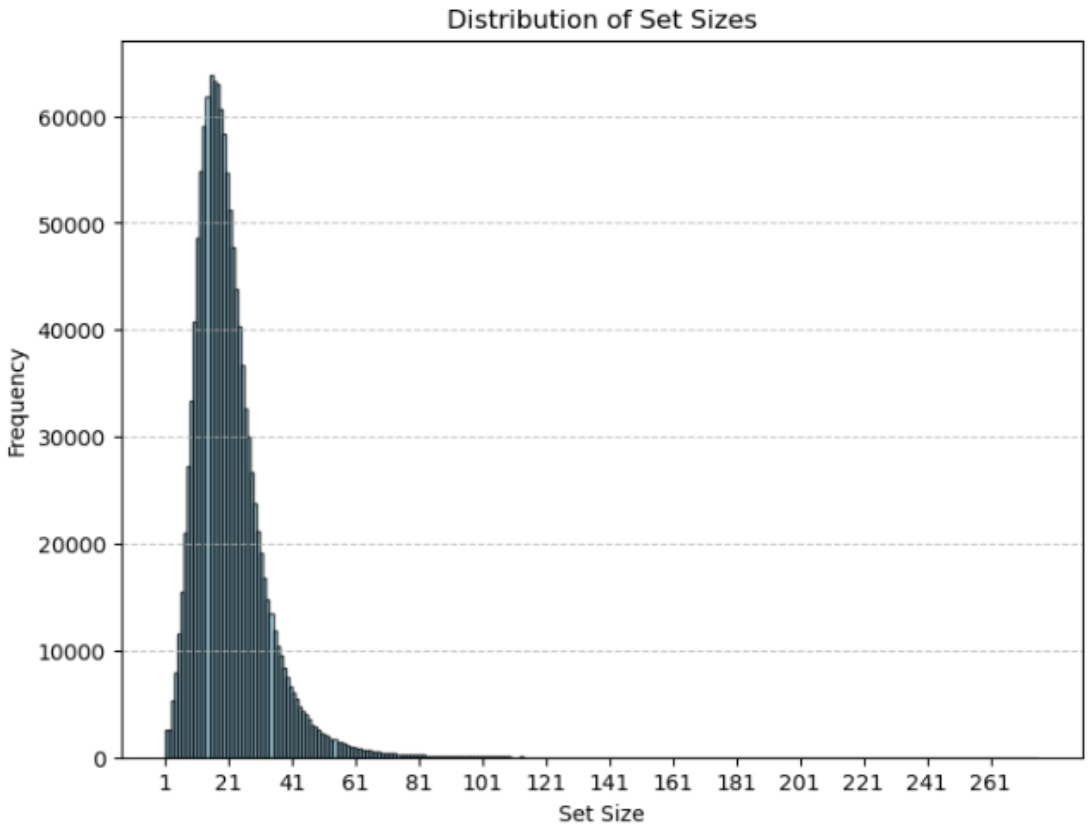

Figure 1

As it can be seen from Figure 1, the set sizes are more concentrated between 15-50 items per each, and the average of the items occurring in each set is 19.

Furthermore, in order to better understand the items, the counts of the most common skills are also plotted on a graph as it can be seen from Figure 2 and as the top 300 most occurring items in the sets are printed to see which ones the most required ones are to get a random job.

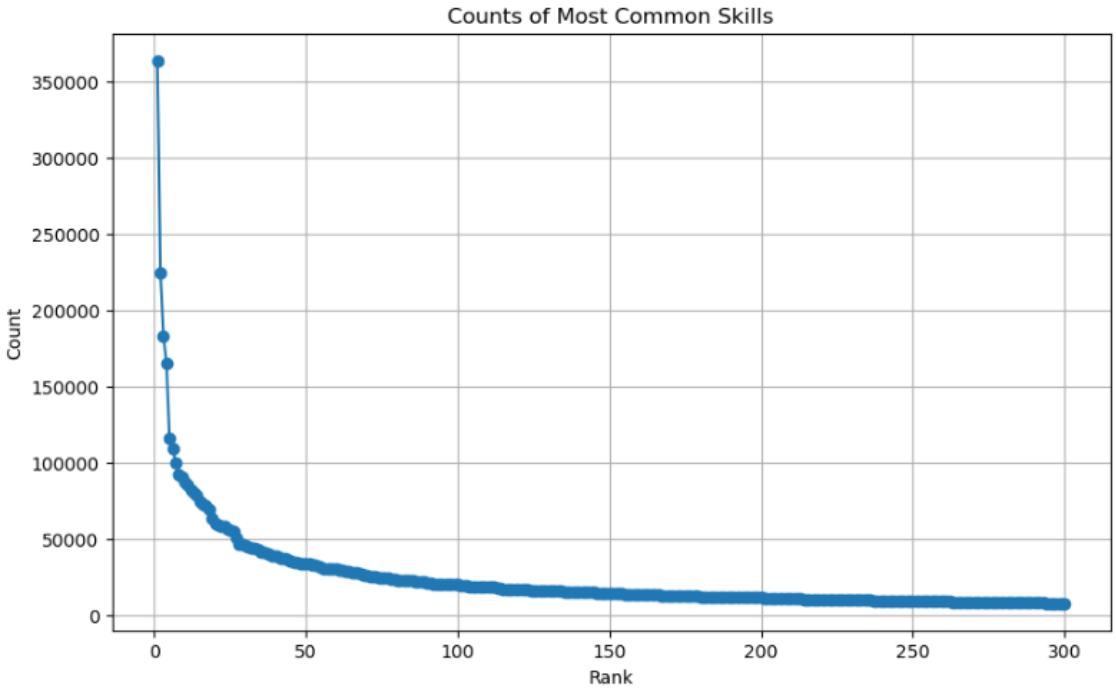

Figure 2

As stated above, among the first 300 occurring's, the top 5 skills were, Communication with 364114 occurring's, Teamwork with 224690 occurring's, Leadership with 182843 occurring's, Customer service 165453 occurring's, Communication skills with 115733 occurring's.

As it can be seen from the skills, there were duplicated values such as Communication and Communication skills since potentially the job posters inserted the skills on their own and didn't choose it automatically from the LinkedIn interface. However, to protect the dataset structure and since there were bunch of the different written versions of the same skills, they are left as they are and the Apriori algorithm were conducted on the original dataset.

3)Experiment and Implementation of Apriori Algorithm

Overview of Apriori Algorithm and Market Basket Analysis

As also indicated in the Introduction part, Market Basket Analysis is an analysis based on the occurrences of item sets together itself in the "baskets" and it is a powerful method to gain deeper understanding of items data patterns. In order a basket to be counted as "frequent" the number of occurring's of the items in the basket must be greater than the minimum support which is determined before initializing the algorithm. For instance, if the minimum support is set as 0.2, in order the itemset to be frequent, the items in this set must occur together in 20% of the whole baskets. Starting with singleton analysis, continuing with doubletons until to the possible number of items that can be put together with regarding the minimum support rule, items are being put together in a basket based on their occurrences and step by step the basket with maximum number of frequent items is achieved.

Another important aspect which needs to be taken into consideration is if an itemset is frequent then its subsets are already frequent. Suppose we have the {Communication, Teamwork, Problem Solving, Customer Service} itemset as a frequent itemset, then we can say that any subset which is the 3-item combination of this frequent itemset is also frequent.

After the achievement of the frequent items, the association rules can be performed based on the confidence threshold to see if we can make a generalization based on what we see in data. The confidence threshold is also determined before applying the Market Basket Analysis and it is generally set as 50%. To give a clear overview of the confidence threshold application to generate association rules, an example can be given. For instance, suppose it is found out that {Communication, Teamwork, Problem Solving, Customer Service} is a frequent itemset and we would like to see if we can generate the following assocation rule: Teamwork + Communication + Problem Solving → Customer Service. This rule indicated the the likelihood of Customer Service can be also added as a skill when we see the

Teamwork, Communication, Problem Solving triplet together. To test this, we must apply to following operation:

$$\frac{\text{Support (Communication, Teamwork, Problem Solving, Customer Service)}}{\text{Support (Communication, Teamwork, Problem Solving)}} > \text{Confidence Threshold}$$

Figure 3

If the support ratio is bigger than the confidence threshold we can associate the Teamwork + Communication + Problem Solving → Customer Service rule.

Figure 4, which is taken from the Mining of Massive Datasets text textbook that summarizes the Apriori algorithm implementation steps is shown below.

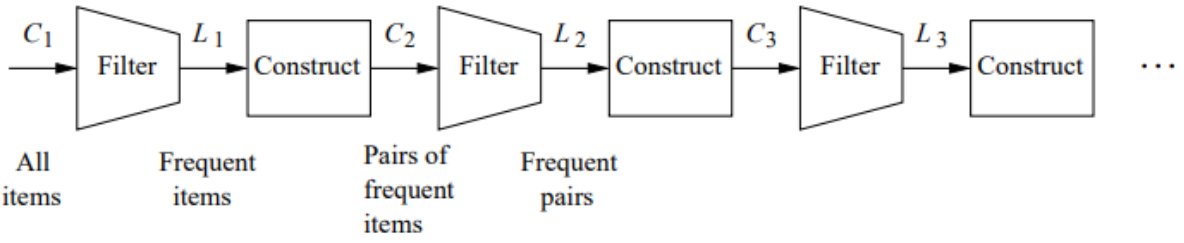

Figure 4

As it can be seen from the provided figure, the implementation of Apriori algorithm consists of iterative steps which includes constructing candidates and filtering them accordingly with the minimum support. C1 corresponds to the list of all items and their count values and then those who meets the minimum support conditions passes through as it can be seen from L1. Then, using the frequent items, frequent pairs are constructed. While construction of the frequent pairs, the items that is focused on are only the frequent items and it creates memory efficiency. However speaking in terms of memory efficiency

based on what is written in the textbook and what have been observed during the implementation, there is a “tyranny” of counting pairs due to having a high main memory requirement. Considering the fact that the dataset has 1.3 million rows meaning that 1.3 million of skillsets, we have also encountered with several time issues during this process since it took 1.5 hours to complete the list of frequent pairs and their counts. Continuing with the C2 step, after obtaining the candidate frequent pairs, the candidate frequent pairs list is also trimmed based on the minimum support threshold and in the L2 step the candidate triplets are constructed. The triplets must also be evaluated due to the minimum support threshold and this iterative process goes like this until the maximum amount of combination of frequent items in a set is reached.

Implementation of Apriori Algorithm

Please note that the Apriori algorithm is implemented from scratch without using any library such as Spark due to having large numbers of issues related to the computer.

Until now, the data exploration and processing step and the general overview of market basket analysis and Apriori algorithm is discussed. In the data preprocessing step, we were left to count the number of items in each skillset. After sorting out the most used 300 skills, it is seen that near 40% of these 300 skills were used less than 10000 times and then it is decided to set the minimum support threshold as 0.01 which also means that we can use 13000 as the threshold. After applying the minimum threshold to the dataset, we were left with 167 frequent skills which are ready to pass to the construction of candidate pairs step. As also indicated above, processing this step took 1.5 hours. The number of candidate pairs found was 27722. After the candidate pairs were constructed, the data frame is extracted as Excel file to see if there are any anomalies, and it is seen that there were several duplicated pairs. Then before removing the duplicated pairs, the candidate pairs were converted to the frequent pairs by removing the pairs which do not satisfy the minimum support threshold. Then again, the

duplicated pairs were checked and found. Consequently, these duplicated pairs were removed from the frequent pairs list, and we were left with 170 frequent pairs.

Secondly, in order to contruct the triplets, the frequent pairs were used. In order to better understand the logic used to construct the triplets also considering eliminating the potential duplication a Figure is provided below.

	Skill1	Skill2	PairCounts
0	Communication	Teamwork	138380
1	Communication	Leadership	116612
2	Communication	Customer service	81236
4	Communication	Customer Service	57077
5	Communication	Problem Solving	66141
...	...	...	...
15052	Standing	Walking	15334
15216	Written Communication	Verbal Communication	15250
16039	Written communication	Verbal communication	14652
17557	Vision	Dental	13502
18386	Dental Insurance	Vision Insurance	15515

170 rows × 3 columns

Figure 5

As provided in Figure 5, a snapshot of the frequent pairs can be seen. To construct the triplets, first the pair of the row is considered as 1st and 2nd elements of the triplet and the second skill is considered as the third skill of the candidate triplet. For instance, {Communication, Teamwork, Leadership} is the first candidate triplet. In order to construct the second candidate triplet, again the skills of first row were considered as the first and second elemnts of the candidate triplet but this time Skill 2 of the third

row is taken as the third element of the potential triplet {Communication, Teamwork, Customer service}. This iteration goes until Skill 1 of the row that we take its Skill 2 as the third skill for our candidate triplet changes. Then we start to take the second rows 2 skills as the 2 skill of the candidate triplet and add the third rows Skill 2 as Skill 3 to the candidate triplets. This process again iterates over until the Skill 1 of the row that we take its Skill as the third skill for the candidate triplet changes. After constructing every candidate triplet with let's say with "Communication", this iterative step starts for the new skill of the Skill1 column of the frequent pairs dataframe. Let's say after the pairs which has Communication as Skill 1 (by the way the skill who has individually more counts appears as Skill 1 in the frequent pairs due to how we constructed the pairs), the next skill that appears as Skill1 in the frequent pairs is Teamwork. Then the same iteration process starts for Teamwork (since Communication is individually more frequent than Teamwork variable it never appears as Skill 2), and the algortihm iterates over frequent pairs which has Teamwork as Skill1.

With using the logic stated above, candidate triplets were constructed. After counting of occurence of triplets together in each set and by trimming them using the minimum support threshold we were left with 60 triplets which each triplet has more occurence than 13000. Using the frequent triplets, candidate fourlets were constructed using a similar logic above provided but this time we take Skill2 of the triplet as the row changer condition. It makes sense to use Skill 2 in this case since, also when the Skill1 changes, Skill2 also changes so we capture every potential candidate fourlets. The number of obtained potential fourlets were 214 and again, we apply the minimum support threshold to the fourlets and we were left only with 8 frequent fourlets.

Furthermore, to see whether if we can proceed with any fivelets, we also constructed fivelets and obtained their counts together, however neither of them passed through the minimum threshold so the frequent itemsets with maximum number of elements in it are fourlets. The frequent items can be seen below as Figure 6.

Skill1	Skill2	Skill3	Skill4	Count
Communication	Teamwork	Leadership	Customer service	13739
Communication	Teamwork	Leadership	Problem Solving	14257
Communication	Teamwork	Leadership	Problemsolving	15533
Communication	Teamwork	Customer service	Problemsolving	21298
Communication	Teamwork	Customer service	Attention to detail	13094
Communication	Teamwork	Customer service	Time management	13086
Communication	Teamwork	Problemsolving	Attention to detail	14234
Communication	Teamwork	Problemsolving	Time management	14976

Figure 6

As it can be seen from the above provided Figure 6, the “most” frequent itemset is {Communication, Teamwork, Customer service, Problemsolving} with 21298 occurences. A association rule study is conducted based on the mentioned itemset considering all of the combinations of association rules that can be obtained from this itemset. The confidence threshold to construct the association rule were considered as 0.5.

The first rule was (Communication - Team - Customer Service -> Problemsolving) measured by the association rule formula provided as an example in Figure 3. In order to validate this relationship if it is associated or not, first the support of the (Communication, Teamwork, Customer service, Problemsolving) must be calculated. Then the support of the left hand side values together must be calculated (Communication, Teamwork, Customer service).

$$\frac{\text{Support (Communication, Teamwork, Problem Solving, Customer service)}}{\text{Support (Communication, Teamwork, Customer service)}} > \text{Confidence Threshold}$$

$$0.5$$

Figure 7

As we already obtained previously, the support of (Communication, Teamwork, Customer service, Problemsolving) is 21298 and the support of (Communication, Teamwork, Customer service) is 40032. When we do the division, we get the value 0.53 which is higher than the confidence threshold ratio so we can say that this is association is valid.

Applying the same method for every possible rule combinations of the “most” frequent itemset, we get tested 14 potential association rules. The results are provided above.

Communication - Team - Customer Service -> Problemsolving is a valid association rule

0.5320243804956035

Communication - Team - Problemsolving -> Customer service isa valid association rule

0.5441770146660535

Communication - Problemsolving - Customer service -> Teamwork isa valid association rule

0.6486371250190346

Teamwork - Problemsolving - Customer service -> Communication isa valid association rule

0.8496768531077954

Communication - Teamwork-> Customer service - Problemsolving is a not valid association rule

0.1539095244977598

Communication - Customer service-> Teamwork - Problemsolving is a not valid association rule

0.2621744054360136

Communication - Problemsolving-> Teamwork - Customer service is a not valid association rule 0.3255828173966216

Teamwork - Customer service -> Communication - Problemsolving is a not valid association rule 0.36886040872878423

Teamwork - Problemsolving -> Communication - Customer service is a not valid association rule 0.4407516245188527

Customer service - Problemsolving -> Communication - Teamwork is a valid association rule 0.5139974901052226

Problemsolving -> Communication - Teamwork - Customer service is a not valid association rule 0.23450265354209332

Customer service -> Teamwork - Problemsolving - Communication is a not valid association rule 0.1287253782040821

Teamwork -> Problemsolving - Communication - Customer service is a not valid association rule 0.09478837509457474

Communication -> Teamwork - Problemsolving - Customer service is a not valid association rule 0.0584926698781151

Additional Analysis

In order to validate our findings, the corresponding links of our “most” frequent itemset {Communication, Teamwork, Customer service, Problemsolving} were extracted as Excel file. Then, since the links contains the job names, using the “text to columns” feature of Excel, the job titles of the most frequent itemset were obtained. This data is uploaded to Jupyter Notebook and a word count

analysis is applied to this dataset to see whether if we can find some useful insights from it. After sorting out the top 50 keywords in the jobs list, it is seen that Communication, Teamwork, Customer service and Problemsolving skills are mostly required in the manager, customer service, supervisor, assistant, representative, sales,associate, retail, advisor, technician, leader, speacialist, senior, consultant, maintenance and operations positions. In other words, in order to get hired in these positions you should have good communication, teamwork, customer service and problem solving skills.

4)Discussion

As a summary, we have applied Apriori algortihm which is a Market Basket Analysis technique to job skills post in LinkedIn from scratch. We have found out 8 4-item frequent sets among 1.3 million and we couldn't find frequent itemsets with 5 members which seemed quite interesting since we have 1.3 million sets and the average number of skill in each set is 19. Furthermore, during the creating of pairs, triplets and fourlets we have seen that filtering before before processing, has a good impact on effective memory usage. We again applied association rule technqiue to our most frequent itemset to see if we can generate some rules and in fact we came up with several rules. Lastly, we conducted an additional analysis by extracting the job titles in the job_links and by using most occured keywords in the job titles, we have found out that the jobs which requires Communication, Teamwork, Customer service and Problemsolving skills.

5)Declaration&Resources

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.“

Resources:

-Data Source: https://www.kaggle.com/datasets/asaniczka/1-3m-linkedin-jobs-and-skills-2024

- Mining of Massive Datasets (A. Rajaraman, J.Ullman)

-Link to Google Colab:

https://colab.research.google.com/drive/1VBN9MoT318X_LlOUNyaHwgcG3hwURS-_?usp=sharing