INNOMATICS RESEARCH LABS

Market Basket Analysis

Unorganized Transactional Data

Data Processed by the Algorithm

Intelligent Associations

1

2

3

Market Basket Motivation

INNOMATICS RESEARCH LABS

ARM is one of those bread and butter tools that has found widespread use in machine learning, AI etc. Now, in business analytics we typically refer to it as MBA because typically applied in studying customer transactions, trying to figure out what items the customers will be interested in purchasing all based on what items are appearing in their shopping carts.

We have all seen this real time right ? When we go to amazon and search for any item, you would see "frequently brought together" , "customers who bought this item also bought". How does amazon know/ suggest these items ?

That's what we are going to learn today !

where did these suggestions come from ?

What combination of items typically appear in the carts together, and how is this knowledge useful?

“Frequently Bought Together” → Association
“Customers who bought this item also bought” → Recommendation

Contents

INNOMATICS RESEARCH LABS

- Market Basket Analysis
- Association rule mining
- Apriori Algorithm
- Example

Motivation behind uncovering MBA & ARM

When we go grocery shopping, we often have a standard list of things to buy. Each shopper has a distinctive list, depending on one's needs and preferences. A housewife might buy healthy ingredients for a family dinner, while a bachelor might buy beer and chips. Understanding these buying patterns can help to increase sales in several ways. If there is a pair of items, X and Y, that are frequently bought together:

Both X and Y can be placed on the same shelf, so that buyers of one item would be prompted to buy the other.
Promotional discounts could be applied to just one out of the two items.
Advertisements on X could be targeted at buyers who purchase Y.
X and Y could be combined into a new product, such as having Y in flavors of X.
While we may know that certain items are frequently bought together, the question is, how do we uncover these associations?

Product placement in Tesco, UK.

Clustering vs. Association

By definition, clustering is grouping a set of objects in such a manner that objects in the same group are more similar than to those object belonging to other groups.
Whereas, association rules is about finding associations amongst items within large commercial databases.

Clustering is about the data points,
ARM is about finding relationships between the attributes of those data points.

Clustering has to do with identifying similar cases in a dataset (i.e. you want to group your rows).
On the other hand, association has to do with identifying similar dimensions in a dataset (i.e. you want to group your columns).

Example

Now, let's take an example. Suppose we have data on trips and corresponding product purchases as below:

trip id	P1	P2	P3	P4	P5	P6	P7	P8	P9
t1	0	1	1	1	0	0	0	1	0
t2	1	1	1	0	0	0	0	0	0
t3	0	0	0	1	1	1	1	0	0
t4	0	0	1	1	1	1	0	0	0
t5	1	1	0	0	0	0	1	1	1
t6	1	1	1	0	0	0	0	1	1

Where, "1" means purchase and "0" means no-purchase.

Now, let's ask ourselves 2 business questions:
i) Which all trips has similar product purchases?
ii) Which products could be grouped together?

Question (i) would be answered by clustering – where we will look at similarities between trips (ti, tj) based on purchased product dimensions.
Question (ii) would be answered by association rules – where we will look at co-occurrences of products (Pi, Pj) within trips and association rules will be derived based on popular metrics, e.g. support, confidence, lift etc.
a set of text documents are a dataset, and clustering has to do with identifying similar text documents (cases of documents); while association has to do with identifying common themes that co-occur frequently across the set of documents in the dataset.

Applications of Market Basket (Other than Retail segment)

INNOMATICS RESEARCH LABS

- Medical diagnosis - each patient is a transaction and symptoms are "items in the cart". What combination of symptoms do sufferers of a certain illness have in common? Could use predictive analytics once established to help do diagnosis given symptoms.
- Census - each hools, drug abuse, racial m teristics are often found to
- Customer profi s they shop at are items in th gether? (e.g., Walmart / Food City, or Target / Whole Foods)
- Autocomplete - each phrase is a transaction and the words they contain are items in the cart. What combinations of words are frequently found together?

What is probability that a "cart" that has items:
- {Sore Throat}
- {Cough}
- {Temperature > 101.5}
also has {COVID-19}?

Reference Courtesy:
Adam Petrie

Market Basket Analysis

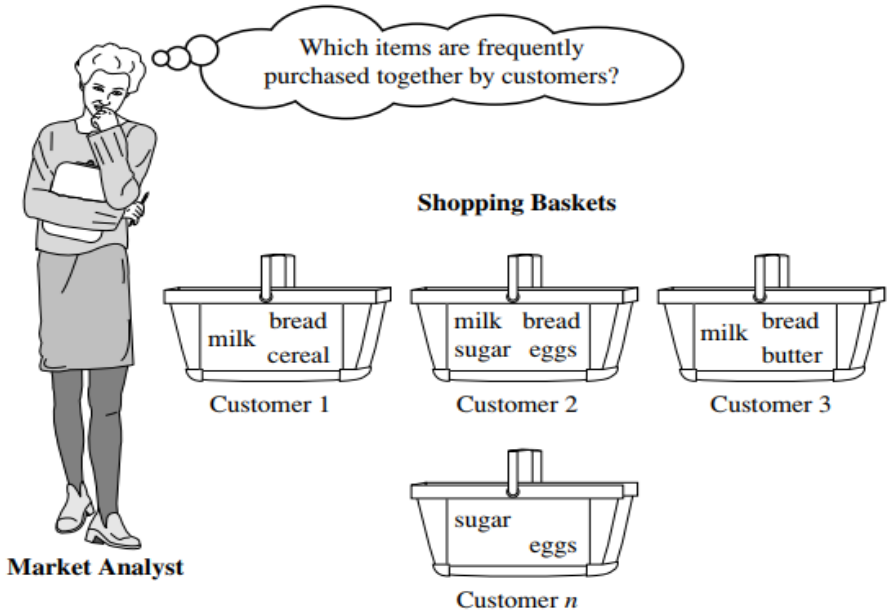
INNOMATICS RESEARCH LABS

- Market Basket Analysis is one of the key techniques used by large retailers to uncover associations between items.
- It works by looking for combinations of items that occur together frequently in transactions.
- It allows retailers to identify relationships between the items that people buy.
- Helps in right placement of products and offers.

Frequent patterns are patterns (e.g., itemset, subsequence's, or substructures) that appear frequently in a data set.

For example, a set of items, such as milk and bread, that appear frequently together in a transaction data set is a frequent itemset

Eg: Market Basket Analysis-This process analyses customer buying habits by finding associations between the different items that customers place in their “shopping baskets”



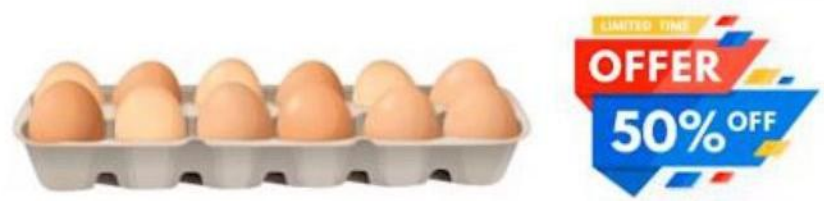
Market Basket Analysis

INNOMATICS RESEARCH LABS

Bread and Jam

Bread and Butter

Laptop and Bag



Market Basket Analysis (Pattern Discovery)

INNOMATICS RESEARCH LABS

A market basket database typically consists of a large number of transaction records.
Each record lists all items purchased during a single customer transaction.
The objective of this data mining exercise is to identify if certain groups of items are usually purchased together. The result is a set of rules, called association rules which summarize item associations as follows:

if [A] is purchased --> then [B] is also purchased, [x%] of time

E.g. - : If age("24..28") and income("50..60K") → buys(" small car")

Mysterious Case of Beer and Diaper

"In the convenience stores , we looked at, on Weekend Nights between 5:00 and 7:00 p.m. , purchase of Beer and purchases of Diaper are highly associated"

--- What do we conclude ?
--- What action should be taken ?

What is Market Basket Analysis?

Understanding behavior of shoppers

What items are bought together

What's in each shopping cart/basket?

Basket data consist of collection of transaction date and items bought in a transaction

Itemset

Retail organizations interested in generating qualified decisions and strategy based on analysis of transaction data

what to put on sale, how to place merchandise on shelves for maximizing profit, customer segmentation based on buying pattern

Examples

INNOMATICS RESEARCH LABS

Rule form :
Antecedent → Consequent [Support, Confidence, Lift]

Rule form: LHS → RHS

IF a customer buys diapers, THEN they also buy beer

diapers → beer

"Transactions that purchase bread and butter also purchase milk"

bread ∧ butter → milk

Customers who purchase maintenance agreements are very likely to purchase large appliances

When a new hardware store opens, one of the most commonly sold items is toilet bowl cleaners

What Is Association Rule Mining?

- Association rule mining
  - Finding frequent patterns, associations, correlations, or causal structures among sets of items in transaction databases
  - Understand customer buying habits by finding associations and correlations between the different items that customers place in their "shopping basket"

- Applications
  - Basket data analysis, cross-marketing, catalog design, loss-leader analysis, web log analysis, fraud detection (supervisor->examiner)

Reference Courtesy:
https://paginas.fe.up.pt/~ec/files_1112/week_04_Association.pdf

How can Association Rules be used?

Stories – Beer and Diapers

- Diapers and Beer. Most famous example of market basket analysis for the last few years. If you buy diapers, you tend to buy beer.
  - T. Blischok headed Terradata's Industry Consulting group.
  - K. Heath ran self joins in SQL (1990), trying to find two itemsets that have baby items, which are particularly profitable.
  - Found this pattern in their data of 50 stores/90 day period.
  - Unlikely to be significant, but it's a nice example that explains associations well.

Ronny Kohavi ICML 1998

Probably mom was calling dad at work to buy diapers on way home and he decided to buy a six-pack as well.

The retailer could move diapers and beers to separate places and position high-profit items of interest to young fathers along the path.

Reference Courtesy:
https://paginas.fe.up.pt/~ec/files_1112/week_04_Association.pdf

Intuitive Example

INNOMATICS RESEARCH LABS

- Wal-Mart knows that customers who buy Barbie dolls (it sells one every 20 seconds) have a 60% likelihood of buying one of three types of candy bars. What does Wal-Mart do with information like that?

- By increasing the price of Barbie doll and giving the type of candy bar free, wal-mart can reinforce the buying habits of that particular types of buyer
- Highest margin candy to be placed near dolls.
- Special promotions for Barbie dolls with candy at a slightly higher margin.
- Take a poorly selling product X and incorporate an offer on this which is based on buying Barbie and Candy. If the customer is likely to buy these two products anyway then why not try to increase sales on X?
- Probably they can not only bundle candy of type A with Barbie dolls, but can also introduce new candy of Type N in this bundle while offering discount on whole bundle. As bundle is going to sell because of Barbie dolls & candy of type A, candy of type N can get free ride to customers houses. And with the fact that you like something, if you see it often, Candy of type N can become popular.

Reference Courtesy:
https://paginas.fe.up.pt/~ec/files_1112/week_04_Association.pdf

Association rule mining

INNOMATICS RESEARCH LABS

- Association rule mining is a form of IF-THEN relationship.

- If product A is bought then there is a high chance of product B is also bought together.

- It may be simple when you have few rows of data. But what if we have big data like large retailers have..?

Association rule mining

INNOMATICS RESEARCH LABS

- At a high level, this is a two step process:
  1. Find all frequent item sets
  2. Generate strog association rules from freq. item sets
- Frequent item sets can be identified using any one of the following methods –
  i. Apriori method
  ii. FP-Growth

Association rule mining

INNOMATICS RESEARCH LABS

- Strong association rules can derived, when the rules satisfy minimum:
  i. Support
  ii. Confidence
  iii. Lift
- One common application of these rules is in the domain of **recommender systems**, where customers who purchased item A are recommended item B.

Rule: if items $X_1, X_2, X_3, \ldots, X_n$ are in the cart, then item Y may also be in the cart.

$$\{X_1, X_2, X_3, \ldots, X_n\} \rightarrow \{Y\}$$

- Support and Coverage - How prevalent or widely applicable is this rule? "Too rare - don't care."
  - Support of rule: to what fraction of carts does this rule apply (i.e., what fraction of carts have *all* items referenced in the rule; the X's and Y)?
  - Coverage: fraction of carts that have all items on the left side of rule (i.e. all X's here)
- Confidence - What is the chance that the rule is correct? When a cart has items $X_1, X_2, \ldots, X_n$, what is the chance that it, indeed, also has item Y?
- Lift - Compared to the overall probability that Y is in a cart, how many times more likely is it that Y is in a cart once we know that items X_1, X_2, etc., are in there?

Goal is to find rules with a relatively high support/coverage so that insight derived from them is widely applicable

Confidence is a conditional probability:
P(Y in cart | all Xs in cart)

Lift reveals how much probability of finding Y changes when we know the cart composition

Reference Courtesy:
Adam Petrie

Association rule mining

INNOMATICS RESEARCH LABS

i. **Support** - This is the percentage of orders that contains the item set.

```
support {A, B}  = freq {A, B} / N
```

ii. **Confidence** - Given two items, A and B, confidence measures the percentage of times that item B is purchased, given that item A was purchased. This is expressed as:

```
confidence {A->B} = support{A,B} / support{A}
```

Association rule mining

INNOMATICS RESEARCH LABS

iii. **Lift** - Given two items, A and B, lift indicates whether there is a relationship between A and B, or whether the two items are occuring together in the same orders simply by chance (ie: at random).

- Unlike the confidence metric whose value may vary depending on direction (eg: confidence{A->B} may be different from confidence{B->A}), lift has no direction.

- This means that the lift{A,B} is always equal to the lift{B,A}:

```
lift{A,B} = lift{B,A} = support{A,B} / (support{A} * support{B})
```

Association rule mining

INNOMATICS RESEARCH LABS

iii. Lift

- lift = 1 implies no relationship between A and B.

  (ie: A and B occur together only by chance)

- lift > 1 implies that there is a positive relationship between A and B.

  (ie: A and B occur together more often than random)

- lift < 1 implies that there is a negative relationship between A and B.

  (ie: A and B occur together less often than random)

Support: It's basically the frequency of items which we have bought and what are the combinations of the frequency of item we have bought. With this, we filter out the items which have been bought less frequently !!

Confidence how often the items say A & B appear together given the number of times A occur

If some one is buying a and b together but not c rule out C at that point !! we obviously do not want to analyse problems that has been bought barely

Set up confidence and support values in the Apriori algorithm

Suppose, even after filtering on confidence & support, we still have 5k rules. we need another metric !!

Lift: strength of any rule

$$lift\{A,B\} = lift\{B,A\} = support\{A,B\} / (support\{A\} * support\{B\})$$

Denominator: independent support values of a and b (independent occurrence value)

Random occurrence vs. association

If denominator is more, occurrence of randomness is more rather than association

Understand Support, Confidence

Rule	Support	Confidence
A ⇒ D	2/5	2/3
C ⇒ A	2/5	2/4
A ⇒ C	2/5	2/3
B & C ⇒ D	1/5	1/3

Understand Support, Confidence

INNOMATICS RESEARCH LABS

Tr1	Shoes, *Socks*, Tie, Belt
Tr2	Shoes, *Socks*, Tie, Belt, Shirt, Hat
Tr3	Shoes, Tie
Tr4	Shoes, *Socks*, Belt

Transaction	Shoes	Socks	Tie	Belt	Shirt	Scarf	Hat
1	1	1	1		0	0	0
2	1	1	1	1	1	0	1
3	1	0	1	0	0	0	0
4	1	1	0	1	0	0	0

Socks => Tie

Support ??

Confidence ??

$$Socks \Rightarrow Tie$$

- Support is 50% (2/4)
- Confidence is 66.67% (2/3)

Association Rule Mining: Example

INNOMATICS RESEARCH LABS

Point of Sale Data

Transaction #	Bag	Blush	Nail Polish	Brushes	Concealer	Eyebrow Pencils	Bronzer
1	No	Yes	Yes	Yes	Yes	No	Yes
2	No	No	Yes	No	Yes	No	Yes
3	No	Yes	No	No	Yes	Yes	Yes
4	No	No	Yes	Yes	Yes	No	Yes
5	No	Yes	No	No	Yes	No	Yes
6	No	No	No	No	Yes	No	No
7	No	Yes	Yes	Yes	Yes	No	Yes
8	No	No	Yes	Yes	No	No	Yes
9	No	No	No	No	Yes	No	No
10	Yes	Yes	Yes	Yes	No	No	No

600 customers in a month buy Nail Polishes
400 customers buy Brushes
300 customers buy Brushes and Nail Polishes
1000 Total Customer Base

Prob(Brushes) = 0.4, Prob(Nail Polishes) = 0.6
Prob (Brushes & Nail Polishes) = 0.3

Rule form :
Antecedent → Consequent [Support, Confidence, Lift]

Support :
Occurrence of Brushes and Nail Polishes / Total No. Of Transactions

Confidence :
Occurrence of Brushes and Nail Polishes / Total No. Of Brushes
i.e. 0.75

Lift :
Confidence of Rule / Prob (Nail Polishes)
i.e. 0.75/0.6 = 1.25

Association Rules
[Brushes] --> [Nail Polish] (confidence: 0.75, lift 1.25)
[Blush] --> [Concealer] (confidence: 0.738)

Interpretation :
A customer who purchases brush is 1.25 times likely to purchase nail polish than a randomly chosen customer

Problem Statement (Motivation behind Apriori)

Given a large transactional database (or data that can be thought of as one), how do we:

Main focus for business analytics practioners:
- set up apriori algorithm
- interpret results
-provide actionable insight

analyze, 416,416,712,497,500 quadruplets of products to analyze, and in general choose(10000,k) k-tuples of products to analyze.

Reference Courtesy:
Adam Petrie

Apriori Algorithm

INNOMATICS RESEARCH LABS

The **Apriori algorithm** is used for mining frequent itemsets and devising association rules from a transactional database. The parameters “support” and “confidence” are used. **Support** refers to items’ frequency of occurrence; **confidence** is a conditional probability.
Items in a transaction form an item set. The algorithm begins by identifying frequent, individual items (items with a frequency greater than or equal to the given support) in the database and continues to extend them to larger, frequent itemsets.

It’s a **bottom-up** approach. We started from every single item in the itemset list. Then, the candidates are generated by self-joining. We extend the length of the itemsets one item at a time. The subset test is performed at each stage and the itemsets that contain infrequent subsets are pruned. We repeat the process until no more successful itemsets can be derived from the data.

Algorithm Steps

INNOMATICS RESEARCH LABS

Algorithm
The following are the main steps of the algorithm:
Calculate the support of item sets (of size k = 1) in the transactional database (note that support is the frequency of occurrence of an itemset). This is called *generating the candidate set*.
Prune the candidate set by eliminating items with a support less than the given threshold.
Join the frequent itemsets to form sets of size k + 1, and repeat the above sets until no more itemsets can be formed. This will happen when the set(s) formed have a support *less than* the given support.

Lk: frequent k-itemset, satisfy minimum support
Ck: candidate k-itemset, possible frequent k-itemsets

Apriori Principle

Any subset of a frequent itemset must be frequent

- A transaction containing {beer, diaper, nuts} also contains {beer, diaper}
- {beer, diaper, nuts} is frequent → {beer, diaper} must also be frequent

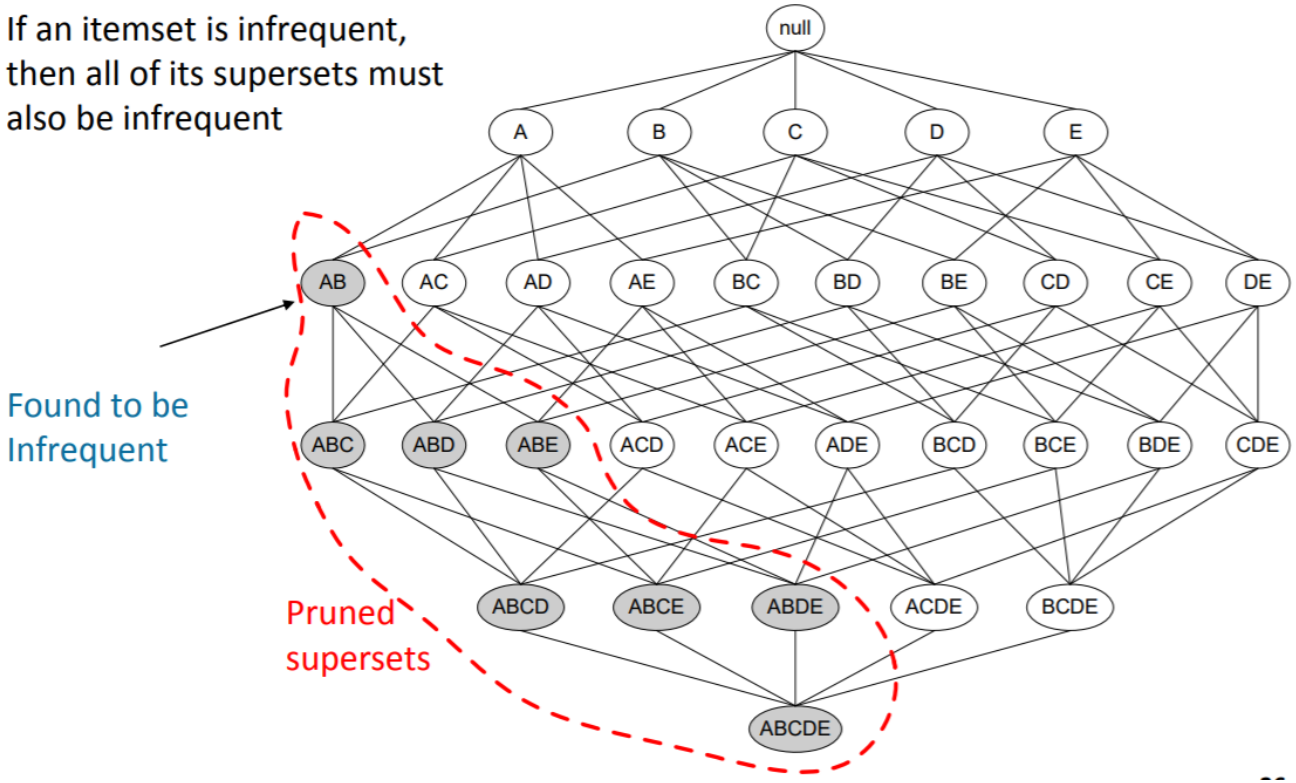
Put simply, the **apriori principle** states that. if an itemset is infrequent, then all its supersets must also be infrequent. This means that if {beer} was found to be infrequent, we can expect {beer, pizza} to be equally or even more infrequent.

- No superset of any infrequent itemset should be generated or tested
  - Many item combinations can be pruned

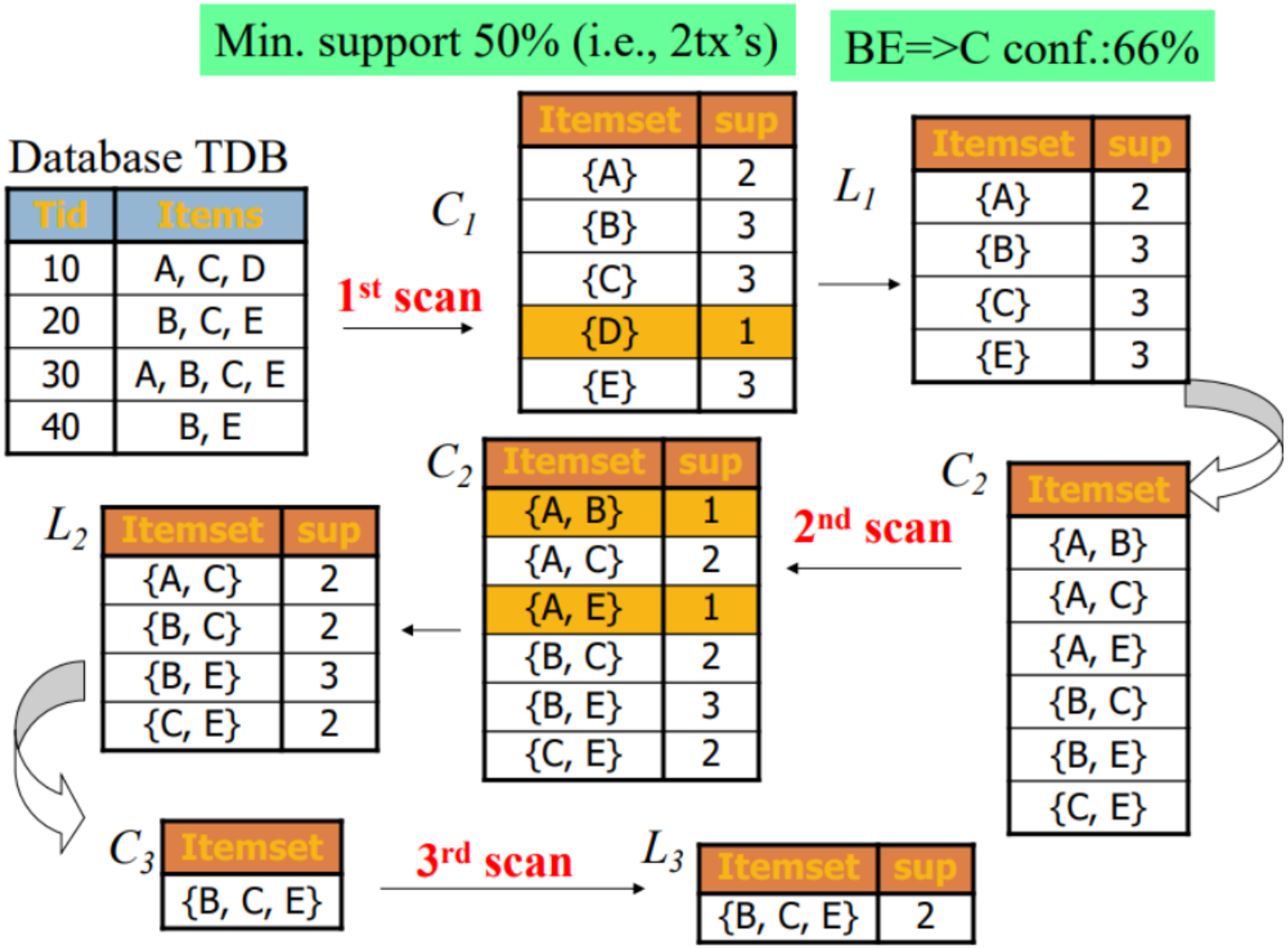
Reference Courtesy:
https://paginas.fe.up.pt/~ec/files_1112/week_04_Association.pdf

Apriori principle for pruning candidates

If an itemset is infrequent, then all of its supersets must also be infrequent



Reference Courtesy:
https://paginas.fe.up.pt/~ec/files_1112/week_04_Association.pdf


Min. support 50% (i.e., 2tx's)
BE=>C conf.:66%
Database TDB
Tid | Items
10 | A, C, D
20 | B, C, E
30 | A, B, C, E
40 | B, E
1st scan
C1
Itemset | sup
{A} | 2
{B} | 3
{C} | 3
{D} | 1
{E} | 3
L1
Itemset | sup
{A} | 2
{B} | 3
{C} | 3
{E} | 3
C2
Itemset
{A, B}
{A, C}
{A, E}
{B, C}
{B, E}
{C, E}
2nd scan
C2
Itemset | sup
{A, B} | 1
{A, C} | 2
{A, E} | 1
{B, C} | 2
{B, E} | 3
{C, E} | 2
L2
Itemset | sup
{A, C} | 2
{B, C} | 2
{B, E} | 3
{C, E} | 2
C3
Itemset
{B, C, E}
3rd scan
L3
Itemset | sup
{B, C, E} | 2

Let’s go over an example to see the algorithm in action. Suppose that the given support is 3 and the required confidence is 80%.

Transaction ID	Items
T1	I1, I2, I3, I4
T2	I2, I3
T3	I3, T4
T4	I2, I3, I4

The transactional database

1-Item set	Support
I1	1
I2	3
I3	4
I4	3

Calculating support for itemsets of size one.

1-Item set	Support
~~I1~~	~~1~~
I2	3
I3	4
I4	3

{I1} 's support is less than the given threshold, so eliminating it.

2-Item sets	Support
I2, I3	3
I2, I4	2
I3, I4	3

Calculating itemsets of size two from the frequent itemsets of size one.

2-Item sets	Support
I2, I3	3
~~I2, I4~~	~~2~~
I3, I4	3

{I2, I4} has a support less than the given support level, so eliminating it.

3-Item sets	Support
I2, I3, I4	2

Generating itemsets of size three.

3-Item sets	Support
I2, I3, I4	2

{I2, I3, I4} has a support less than the given support level, so eliminating it. The algorithm terminates because itemsets of size four cannot be generated.

Flow Chart for Apriori