**GRIFFITH COLLEGE DUBLIN**

**DATA MINING ALGORITHMS AND TECHNIQUES**

Aravindan Srinivasan

[aravindan.srinivasan@student.griffith.ie](mailto:aravindan.srinivasan@student.griffith.ie)

2981707

# Q1. By Hand Apriori – 20 marks

1. Illustrate the steps by hand of using the Apriori algorithm on the example in Table A with support threshold *s*=33.33% (2 items) and confidence threshold *c*=60%.
   1. Show the candidate and frequent itemsets for each database scan.
   2. Enumerate all the final frequent itemsets.
   3. Indicate the association rules that are generated
   4. Highlight the strong ones
   5. Sort them by confidence.

|  |  |
| --- | --- |
| Transaction ID | Items |
| T1 | HotDogs, Buns, Ketchup |
| T2 | HotDogs, Buns |
| T3 | HotDogs, Coke, Chips |
| T4 | Chips, Coke |
| T5 | Chips, Ketchup |
| T6 | HotDogs, Coke, Chips |

Table A

***Solution:***

* **Show the candidate and frequent itemsets for each database scan.**

Support threshold =33.33% => threshold is at least 2 transactions.

|  |  |  |  |
| --- | --- | --- | --- |
| Applying Apriori Pass (k) | Candidate k-itemsets and their support | Frequent k-itemsets |  |
| k=1 | HotDogs(4)  Buns(2)  Ketchup(2)  Coke(3)  Chips(4) | HotDogs  Buns  Ketchup  Coke  Chips |  |
| k=2 | {HotDogs, Buns}(2)  {HotDogs, Ketchup}(1)  {HotDogs, Coke}(2)  {HotDogs, Chips}(2)  {Buns,Ketchup}(1)  {Buns,Coke}(0)  {Buns,Chips}(0)  {Ketchup,Coke}(0)  {Ketchup,Chips(1)  {Coke, Chips}(3) | {HotDogs, Buns},  {HotDogs, Coke},  {HotDogs, Chips},  {Coke, Chips} |  |
| k=3 | {HotDogs,Buns,Coke}(0)  {HotDogs,Buns,Chips}(0)  {HotDog,Coke,Chips}(2) | {HotDogs, Coke, Chips} |  |
| k=4 | {HotDog,Coke,Chips} | |  |

|  |
| --- |
| ***Note that***{HotDogs, Buns, Coke} *and* {HotDogs, Buns, Chips} *are not candidates when* k=3 *because their subsets* {Buns, Coke} *and* {Buns, Chips} *are not frequent*. |

* **Enumerate all the final frequent itemsets.**

|  |
| --- |
| All Frequent Itemsets: {HotDogs}, {Buns}, {Ketchup}, {Coke}, {Chips}, {HotDogs, Buns}, {HotDogs, Coke}, {HotDogs, Chips}, {Coke, Chips}, {HotDogs, Coke, Chips}. |

* **Indicate the association rules that are generated**

|  |
| --- |
| Association rules:  {HotDogs, Buns} would generate: HotDogs -> Buns (2/6=0.33, 2/4=0.5) and  **Buns -> HotDogs (2/6=0.33, 2/2=1);**  {HotDogs, Coke} would generate: HotDogs -> Coke (0.33, 0.5) and  **Coke -> HotDogs (2/6=0.33, 2/3=0.66);**  {HotDogs, Chips} would generate: HotDogs -> Chips (0.33, 0.5) and  Chips -> HotDogs (2/6=0.33, 2/4=0.5); `  {Coke, Chips} would generate: **Coke -> Chips (3/6=0.5, 3/3=1)** and  **Chips -> Coke (3/6=0.5, 3/4=0.75);**  {HotDogs, Coke, Chips} would generate: HotDogs -> Coke ^ Chips (2/6=0.33, 2/4=0.5),  **Coke -> Chips ^ HotDogs (2/6=0.33, 2/3=0.66)**,  Chips -> Coke ^ HotDogs (2/6=0.33, 2/4=0.5),  **HotDogs ^ Coke -> Chips(2/6=0.33, 2/2=1)**,  **HotDogs ^ Chips -> Coke(2/6=0.33, 2/2=1)** and  **Coke ^ Chips -> HotDogs(2/6=0.33, 2/3=0.66)**. |

* **Highlight the strong ones**
* **Sort them by confidence**.

With the confidence threshold set to 60%, the Strong Association Rules are (sorted by confidence):

|  |
| --- |
| 1. Coke -> Chips (0.5, 1)   2. Buns -> HotDogs (0.33, 1);  3. HotDogs ^ Coke -> Chips(0.33, 1)  4. HotDogs ^ Chips -> Coke(0.33, 1)  5. Chips -> Coke (0.5, 0.75)  6. Coke -> HotDogs(0.33, 0.66)  7. Coke -> Chips^HotDogs(0.33, 0.66)  8. Coke^Chips->HotDogs(0.33,0.66) |

1. Illustrate the steps by hand of using the Apriori algorithm on the example in Table B with support threshold *s*=30% (6 items) and confidence threshold *c*=75%.
2. Show the candidate and frequent itemsets for each database scan.
3. Enumerate all the final frequent itemsets.
4. Indicate the association rules that are generated
5. Highlight the strong ones
6. Sort them by confidence.

|  |  |
| --- | --- |
| Transaction ID | Items Purchased |
| 1 | A,B,C,D |
| 2 | B,C,D,E,G |
| 3 | A,C,G,H,K |
| 4 | B,C,D,E,K |
| 5 | D,E,F,H,L |
| 6 | A,B,C,D,E,L |
| 7 | A,D,E,F,L |
| 8 | B,I,K,L |
| 9 | C,D,F,L |
| 10 | A,B,D,E,K |
| 11 | C,D,H,I,K |
| 12 | B,C,E,K |
| 13 | B,C,D,F |
| 14 | A,B,C,D |
| 15 | C,H,I,J |
| 16 | A,E,F,H,L |
| 17 | H,K,L |
| 18 | A,B,D,H,K |
| 19 | D,E,K |
| 20 | B,C,D,E,H |

Table B

***Solution:***

* **Show the candidate and frequent itemsets for each database scan.**

support threshold *s*=30% (6 items) and confidence threshold *c*=75%.

|  |  |  |  |
| --- | --- | --- | --- |
| Applying Apriori Pass (k) | Candidate k-itemsets and their support | Frequent k-itemsets |  |
| k=1 | A(8)  B(11)  C(12)  D(14)  E(10)  F(5)  G(2)  H(8)  I(3)  J(1)  K(9)  L(7) | A(8)  B(11)  C(12)  D(14)  E(10)  H(8)  K(9)  L(7) |  |
| k=2 | AB(5)  AC(4)  AD(6)  AE(4)  AH(3)  AK(3)  AL(3)  BC(8)  BD(9)  BE(6)  BH(1)  BK(5)  BL(2)  CD(9)  CE(4)  CH(4)  CK(4)  CL(2)  DE(8)  DH(4)  DK(4)  DL(4)  EH(3)  EK(4)  EL(4)  HK(4)  HL(3)  KL(2) | AD(6)  BC(8)  BD(9)  BE(6)  CD(9)  DE(8) |  |
| k=3 | ACD(3),ADE(3),  ABD(5),BCD(7),  BCE(5),BDE(5),  ACE(1) | BCD(7) |  |

* **Enumerate all the final frequent itemsets.**

|  |
| --- |
| All Frequent Itemsets: A,B,C,D,E,H,K,L,AD,BC,BD,BE,CD,DE,BCD |

* **Indicate the association rules that are generated**

|  |
| --- |
| Association rules:  **{A D} would generate: A -> D (6/20=0.3, 6/8=0.75)** **and**  C -> B (6/20=0.30, 6/12=0.5);  **{B C} would generate: B -> C (8/20=0.4, 8/11=0.72) and**  C -> B (8/20=0.4, 8/12=0.66);  **{B D} would generate: B -> D (9/20=0.45, 9/11=0.81) and**  D -> B (9/20=0.45, 9/14=0.64);  {B E} would generate: B -> E (6/20=0.3, 6/11=0.54)and  E -> B (6/20=0.3, 6/10=0.6)  **{C D} would generate: C -> D (9/20=0.45, 9/12=0.75)** **and**  D -> C (9/20=0.45, 9/14=0.64);  {D E} would generate: D -> E (8/20=0.4, 8/14=0.57) and  **E -> D (8/0.4=0.40, 8/10=0.8);**  **{B, C, D}** would generate: **B^C -> D = (7/20=0.35,7/8=0.87)**  **B^D -> C = (7/20=0.35,7/9=0.77)**  **C^D -> B = (7/20=0.35,7/9=0.77)**  B -> C^D = (7/20=0.35,7/11=0.63  C -> B^D = (7/20=0.35,7/12=0.58  D -> B^C = (7/20=0.35,7/14=0.5 |

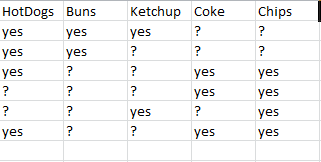
* **Highlight the strong ones**
* **Sort them by confidence**.

With the confidence threshold set to 75%, the Strong Association Rules are (sorted by confidence):

|  |
| --- |
| **B^C -> D = (7/20=0.35,7/8=0.87)**  **B -> D (9/20=0.45, 9/11=0.81)**  **E -> D (8/0.4=0.40, 8/10=0.8)**  **B^D -> C = (7/20=0.35,7/9=0.77)**  **C^D -> B = (7/20=0.35,7/9=0.77)**  **C -> D (9/20=0.45, 9/12=0.75)**  **A -> D (6/20=0.3, 6/8=0.75)** |

# Q2. By Hand FP-Growth – 20 marks

1. Use the transactional database from Table A in the previous exercise with same support threshold and build a frequent pattern tree (FP-Tree).
2. Show the header table.



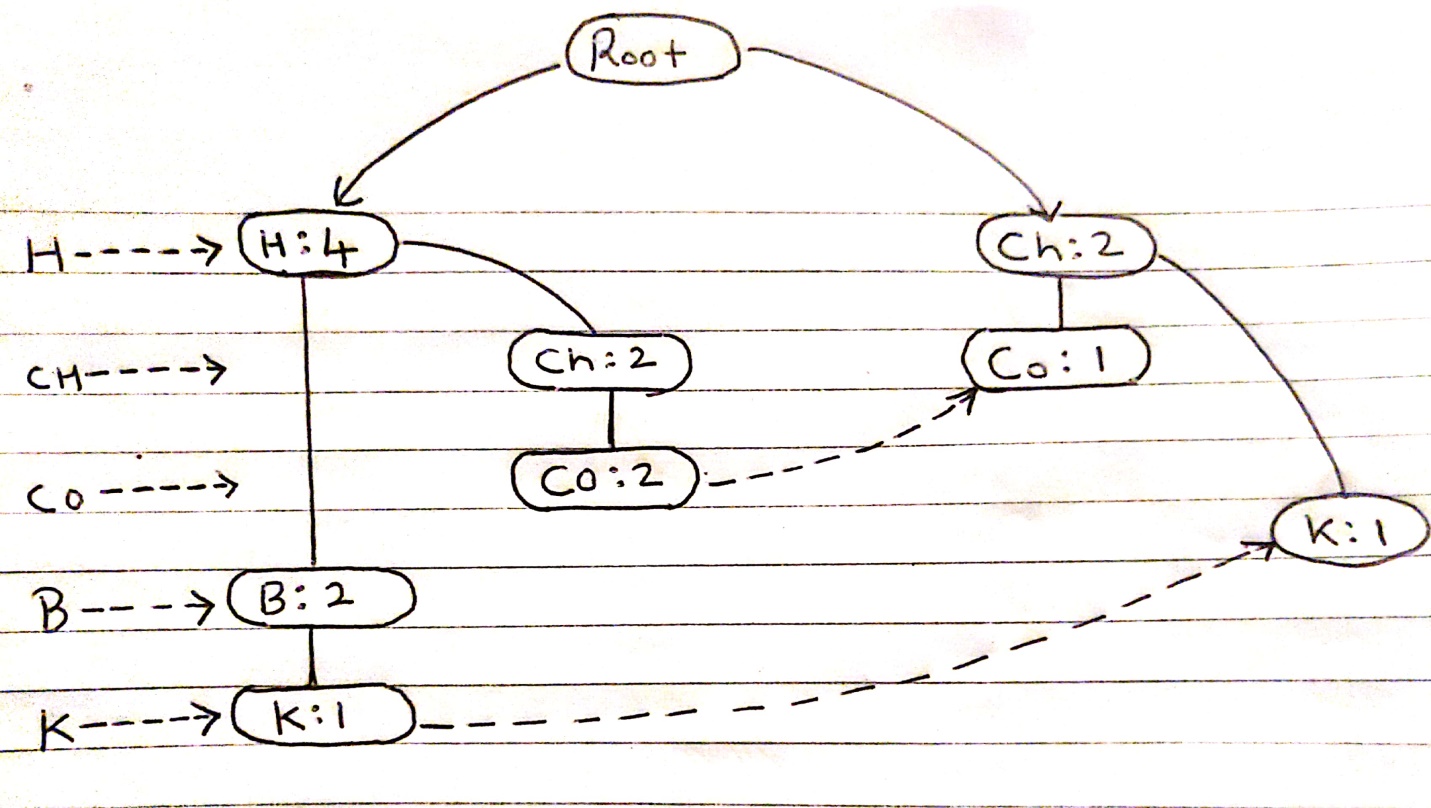
|  |  |  |
| --- | --- | --- |
| Items | Code | Support |
| Hotdog | H | 4 |
| Chips | Ch | 4 |
| Coke | Co | 3 |
| Buns | B | 2 |
| Ketchup | K | 2 |

Ordered items according to support

**H,Ch,Co,B,K**

|  |  |
| --- | --- |
| **Items** | **Ordered Items** |
| H,Ch,Co,B,K | H,B,K |
| H,B | H,B |
| H,Co,Ch | H,Ch,Co |
| Ch,Co | Ch,Co |
| Ch,K | Ch,K |
| H,Co,Ch | H,Ch,Co |

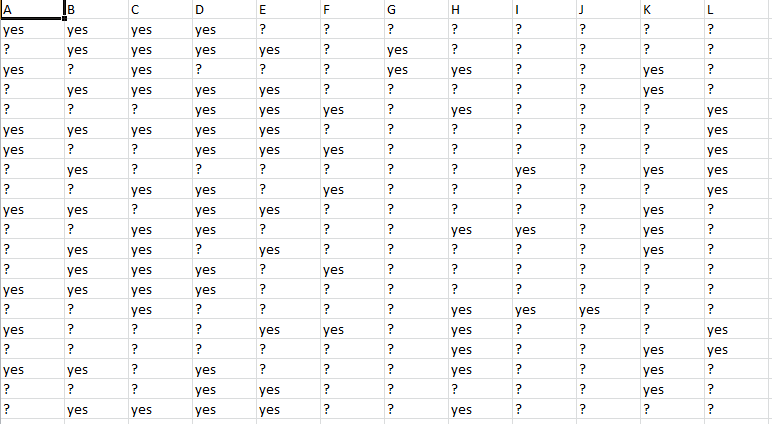
1. Show for each transaction how the tree evolves.



1. List the frequent itemsets.

H,CH,CO are the frequent itemsets

1. Use the transactional database from Table B in the previous exercise with same support threshold and build a frequent pattern tree (FP-Tree).
2. Show the header table.



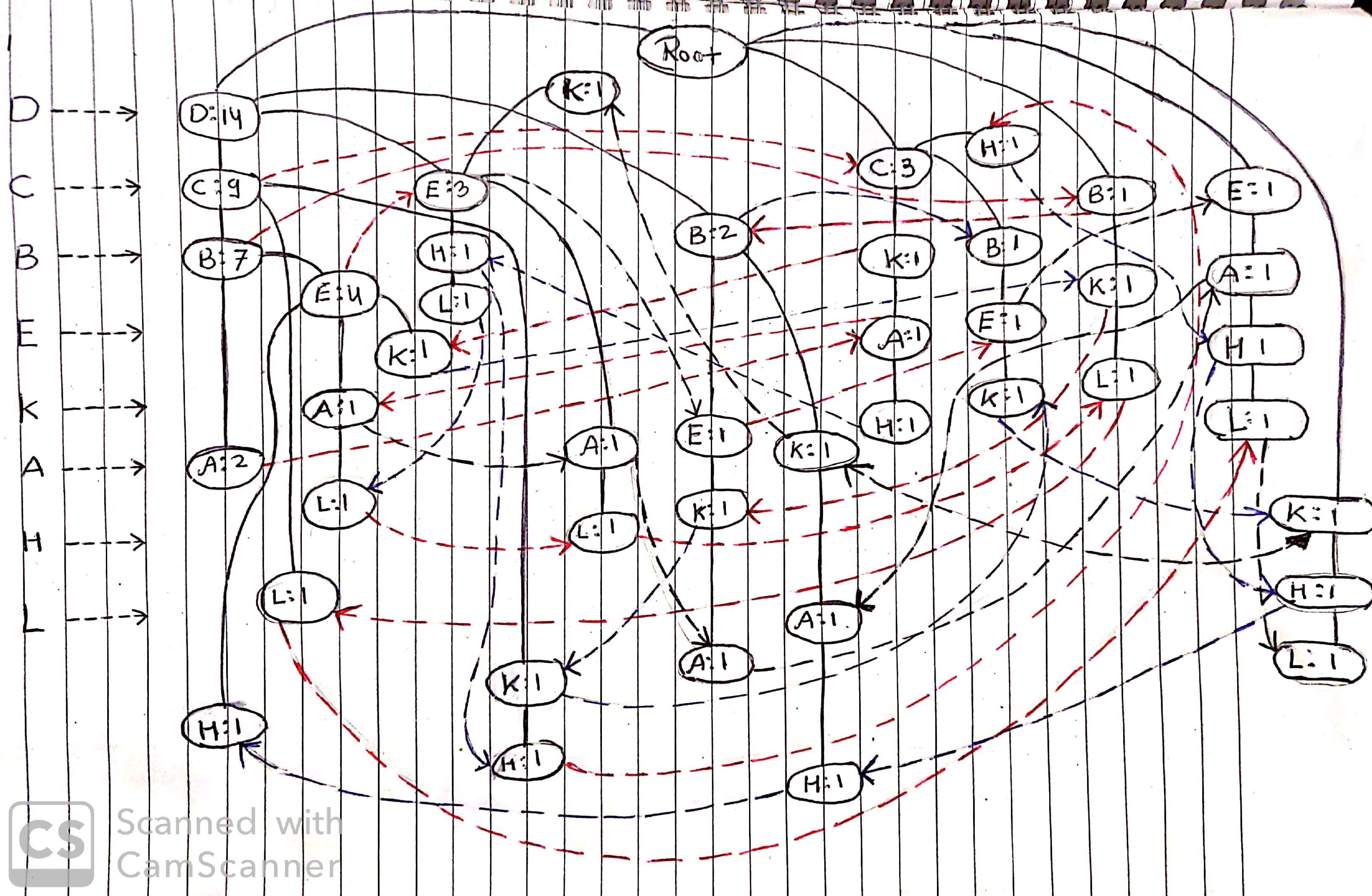
|  |  |
| --- | --- |
| Itemset | Support |
| A | 8 |
| B | 11 |
| C | 12 |
| D | 14 |
| E | 10 |
| F | 5 |
| G | 2 |
| H | 8 |
| I | 3 |
| J | 1 |
| K | 9 |
| L | 7 |

Ordered items according to support

**D,C,B,E,K,A,H,L**

|  |  |  |
| --- | --- | --- |
| Transaction ID | Items Purchased | Orderd item |
| 1 | A,B,C,D | D,C,B,A |
| 2 | B,C,D,E,G | D,C,B,E |
| 3 | A,C,G,H,K | C,K,A,H |
| 4 | B,C,D,E,K | D,C,B,E,K |
| 5 | D,E,F,H,L | D,E,H,L |
| 6 | A,B,C,D,E,L | D,C,B,E,A,L |
| 7 | A,D,E,F,L | D,E,A,L |
| 8 | B,I,K,L | B,K,L |
| 9 | C,D,F,L | D,C,L |
| 10 | A,B,D,E,K | D,B,E,K,A |
| 11 | C,D,H,I,K | D,C,K,H |
| 12 | B,C,E,K | C,B,E,K |
| 13 | B,C,D,F | D,C,B |
| 14 | A,B,C,D | D,C,B,A |
| 15 | C,H,I,J | C,H |
| 16 | A,E,F,H,L | E,A,H,L |
| 17 | H,K,L | K,H,L |
| 18 | A,B,D,H,K | D,B,K,A,H |
| 19 | D,E,K | D,E,K |
| 20 | B,C,D,E,H | D,C,B,E,H |

1. Show how the tree evolves.



1. List the frequent itemsets

# Q3. Weka – Aprior & FP-Growth – 20 marks

Given the following database in Table C with 5 transactions and a minimum support threshold of 60% and a minimum confidence threshold of 80%. **N.B.** you will have to create the csv file for Table C. Please include this in your submission.

1. Using Apriori with Weka show
2. Itemsets (screen shot)
3. Association Rules (screen shot)
4. Using FP-Growth with Weka show
5. Header table (screen shot)
6. Association Rules (screen shot)
7. A list all strong association rules that contain “A” in the antecedent (Constraint).

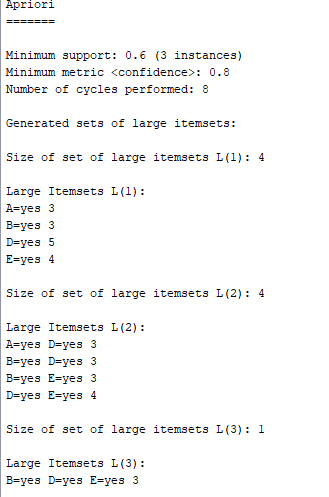
Unfortunately Weka does not provide FP-Tree visualization

|  |  |
| --- | --- |
| **TID** | **Transaction** |
| T1 | A, B, C, D, E, F |
| T2 | B, C, D, E, F, G |
| T3 | A, D, E, H |
| T4 | A, D, F, I, J |
| T5 | B, D, E, K |

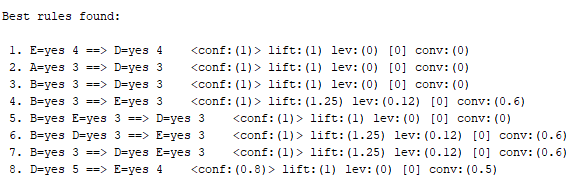
Table C

**a)**

**i)** **Itemsets:-**

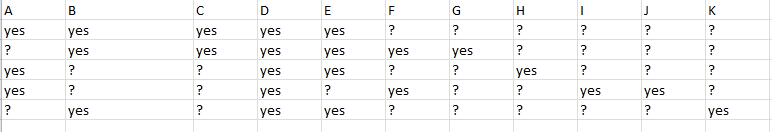


1. **Association Rules:-**

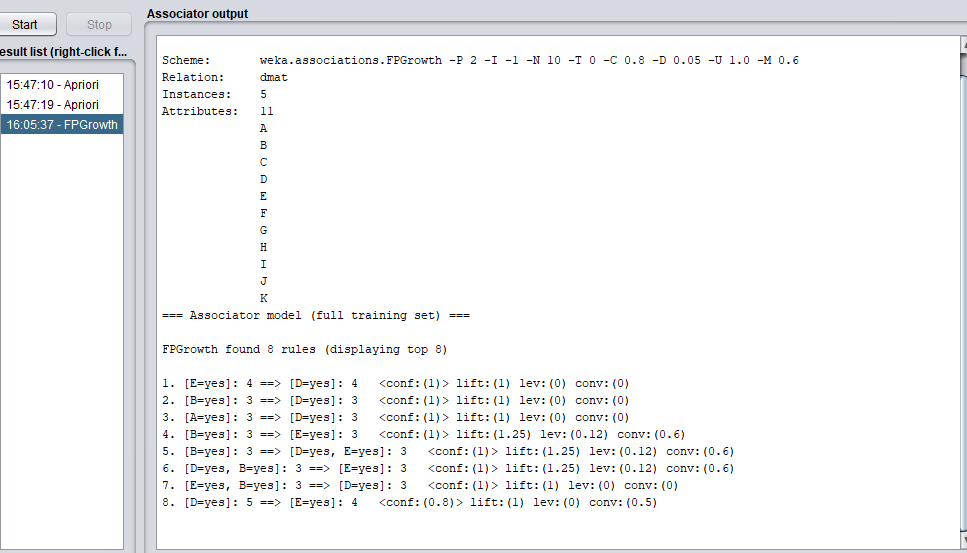


**b)**

**i)** **Header table:-**



**ii) Association Rules:-**



**iii).A list all strong association rules that contain “A” in the antecedent (Constraint).**



As we can see that there is only one strong association that contain “A” in the antecedent(Constraint).

# Q4. Comparison – 20 mark

Using the lecture materials to support your answer, compare the efficiency of both processes.

The difference between these algorithms is how they generate the output. The result is the same, but the process to obtain the result is different. Briefly, Apriori utilize a level-wise approach where it will generate patterns containing 1 items, then 2 items, 3 items, etc. Moreover, it will repeatedly scan the database to count the support of each pattern. On the other hand, FPGrowth utilizes a depth-first search instead of a breadth first search and uses a pattern-growth approach (this means that unlike Apriori, it only considers patterns actually existing in the database.Depending on the number of different items while FP growth runtime increases linearly,depending on the number of transaction and items

# Q5. Apriori investigation of e-commerce data – 20 marks

For this problem you will use some preprocessed and aggregated clickstream data from a real e-commerce site, and use association rule mining to perform market basket analysis on the visitor session data.

There are two primary types of products sold through the above site

1. leg care products,
2. leg ware products.

Each category includes various subcategories and individual products from multiple vendors. There is also a separate categorization of products by specialized "Collections" and by "Assortments." The data collection mechanism, in addition to capturing clickstream page-level data, also captures the information on categories, subcategories, assortments, and collections of products accessed in a given session.

For simplicity, the provided data combines and aggregates visited pages from the log files, category and subcategory names, and product related content pages/categories. The aggregate data contains a total of 182 attributes corresponding to pages or categories. These attributes are listed in the file **Leg-Pages.txt**. The session data is provides in ARFF format in the file**legs.arff** (in a Zip archive). This data contains a total of 7296 sessions (each row in the data). For the purpose of market basket analysis in WEKA, the session data is represented in relational format with unary categorical attributes (a value of "Y" indicates that the corresponding page/category was visited in the session, while a value of "?" indicates that the page/category is missing from the session). Thus, a typical association rule might look similar to the following:

**/Products/Legwear=Y /Products/Legwear/Berkshire=Y ==> Collection: Better Than Bare - Queen=Y**

or

**Category: Health Supplements=Y ==> Subcategory: Bones & Joints=Y**

**a)**

1. **List the top 3 with the number of visits and attribute number**.

* Attribute number 12 :- Total number of visit is 6092
* Attribute number 98 :- Total number of visit is 2819
* Attribute number 100 :- Total number of visit is 2582

These top three attributes are mostly visited

1. **List the bottom 3 with the number of visits and attribute number.**

* Attribute number 74 :- Total number of visit is 31
* Attribute number 01 :- Total number of visit is 32
* Attribute number 54 :- Total number of visit is 32

These three number of attribute are bottom three with the number of visits

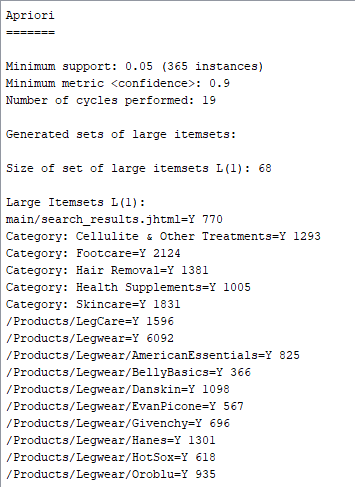
**b)**

Perform association rule mining using Apriori algorithm with the following parameters

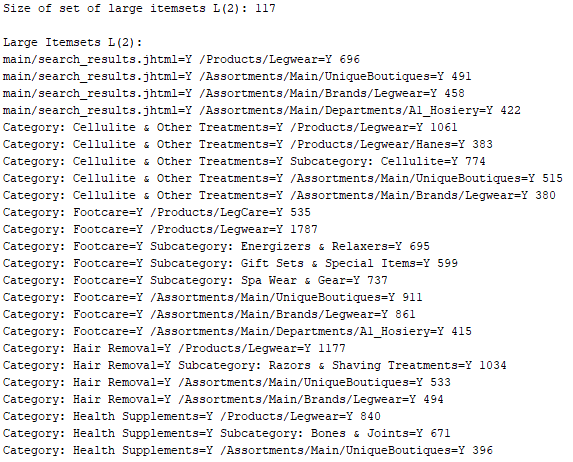
* + "lowerBoundMinSupport" of 0.05
  + confidence O.9 as the minMetric for filtering the rules.
  + set "outputItemsets" to "True" so that you can also view the frequent items sets of different sizes in addition to the rules.
    1. Show the Itemsets (screen shot)
    2. Show the Association Rules (screen shot)

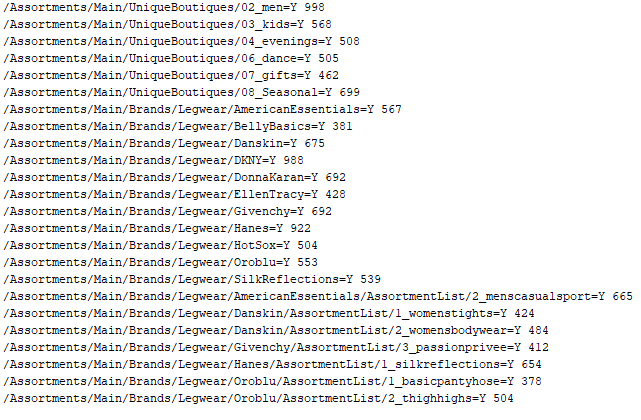
Write a short summary of your observations, including any significant or interesting (e.g., unobvious or unexpected) associations you observe in the data based on the results

1. Show the Itemsets

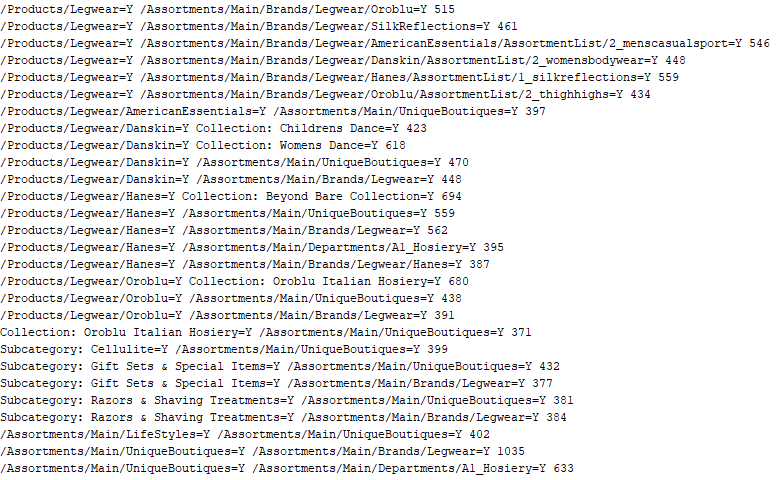


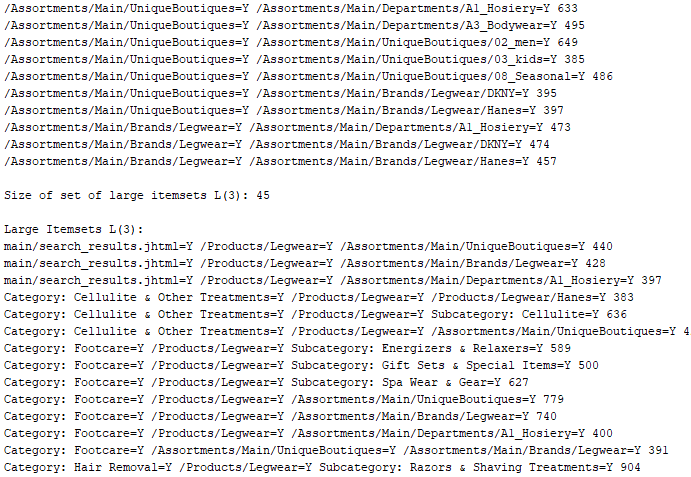


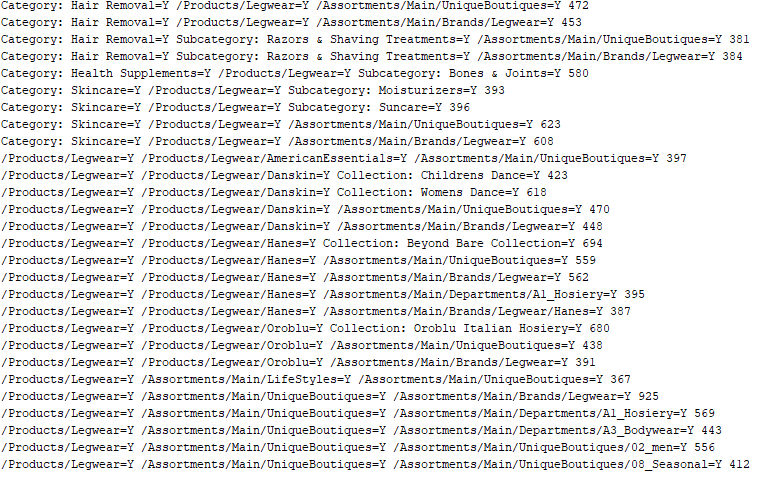






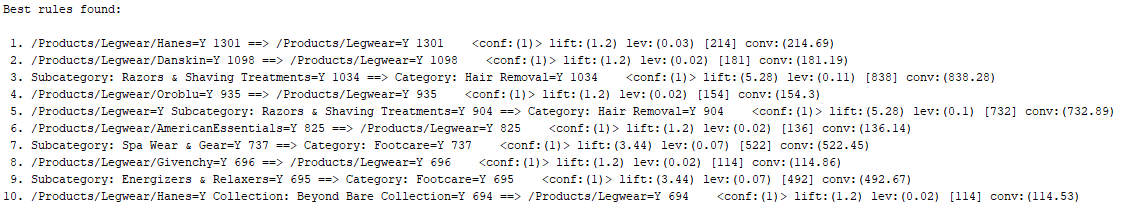




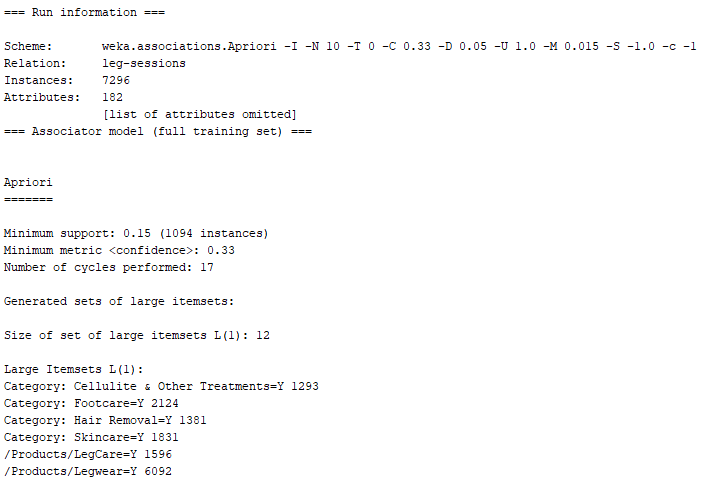


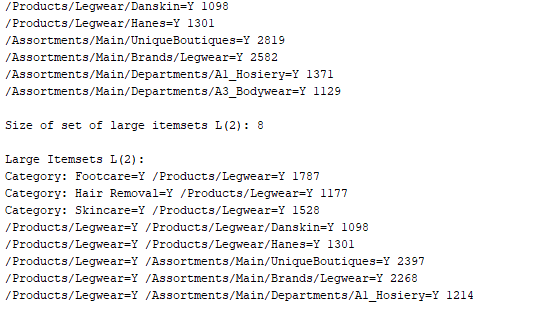


1. **Show the Association Rules**

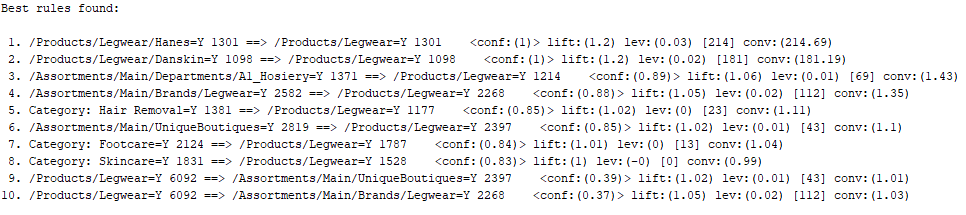


1. Next, run the Apriori algorithm with a lower "lowerBoundMinSupport" so that you can identify associations at a more granular level (e.g., the level of individual products brands rather than higher level categories). You might want to start with 0.025 and go lower if necessary. Experiment with this threshold, as well as the Confidence metrics in different runs and pick the result set that seems to provide the most useful information (e.g., not too many obvious or noisy rules and not too few general rules).
   * **Show the itemsets**





* + **Show the Association Rules**



# Q6. Association Rules investigation of e-commerce data – 20 marks

The marketing department of a financial firm keeps records on customers, including demographic information and the type of accounts. When launching a new product, such as a "Personal Equity Plan" (PEP), a direct mail piece, advertising the product, is sent to existing customers, and a record kept as to whether that customer responded and bought the product. Based on this store of prior experience, the managers decide to use data mining techniques to build customer profile models. In this particular problem we are interested only in deriving (quantitative) association rules from the data (in a future assignment we will consider the use of classification with this data).

The data contains the following fields

|  |  |
| --- | --- |
| **id** | a unique identification number |
| **age** | age of customer in years (numeric) |
| **sex** | MALE / FEMALE |
| **region** | inner\_city/rural/suburban/town |
| **income** | income of customer (numeric) |
| **married** | is the customer married (YES/NO) |
| **children** | number of children (numeric) |
| **car** | does the customer own a car (YES/NO) |
| **save\_acct** | does the customer have a saving account (YES/NO) |
| **current\_acct** | does the customer have a current account (YES/NO) |
| **mortgage** | does the customer have a mortgage (YES/NO) |
| **pep** | did the customer buy a PEP (Personal Equity Plan) after the last mailing (YES/NO) |

Table D

The data is contained in the file**bank-data.csv**. Each record is a customer description where the "pep" field indicates whether or not that customer bought a PEP after the last mailing.

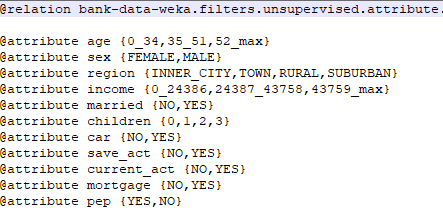
Your goal is to perform Association Rule discovery on the data set using the **Weka** package.

**Note**: Association rule mining requires discretization of continuous variables. This task can be performed in the data transformation step or (in some cases) by the mining program. WEKA is a full data mining suite which includes various preprocessing modules (filters). When using WEKA, you will first apply the relevant preprocessing filters to transform the data before you performing association rule discovery.

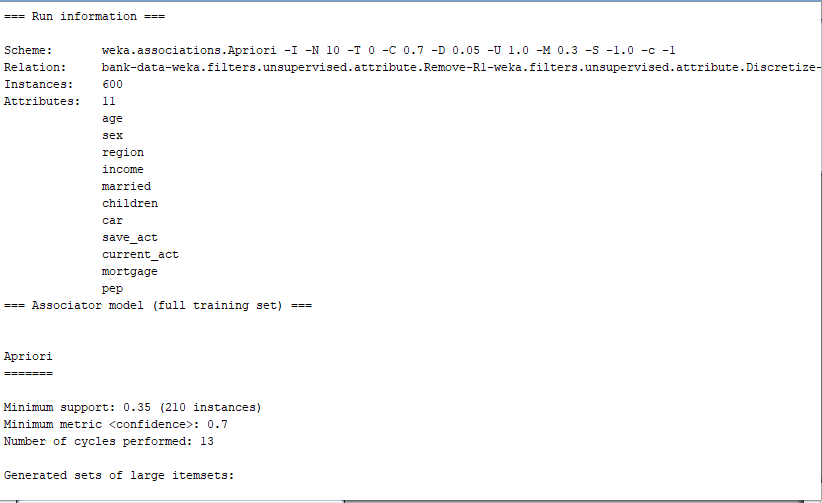
1. First perform the necessary preprocessing steps required for association rule mining. Specifically, the "id" field will need to be removed and the numerical attributes must be discretized. Illustrate this using screen shots.
2. Next perform association rule discovery on the transformed data. In WEKA Apriori algorithm interface set "outputItemsets" to "True" so that you can also view the frequent items sets of different sizes in addition to the rules. Illustrate this using a screen shot.
3. Experiment with different parameters so that you get at least 20-30 strong rules (e.g., rules with high confidence which at the same time have relatively good support). Illustrate this using a screen shot.
4. List the top 5 most "interesting" rules and for each specify the following:
   * an explanation of the pattern and why you believe it is interesting based on the business objectives of the company;
   * any recommendations based on the discovered rule that might help the company to better understand behavior of its customers or in its marketing campaign.

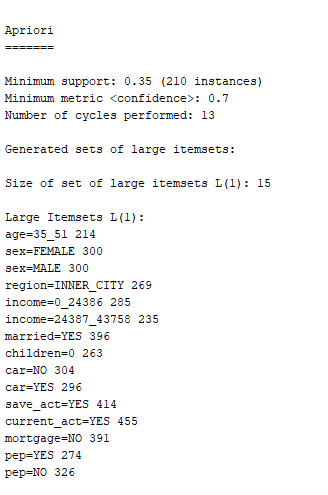
**Note**: The top 5 most interesting rules are most likely not the top 5 in the result set of the Apriori algorithm. They are rules that, in addition to having high support and confidence, also provide some non-trivial, actionable knowledge based on the underlying business objectives.

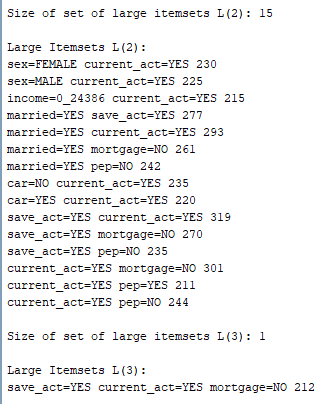
1. First perform the necessary preprocessing steps required for association rule mining. Specifically, the "id" field will need to be removed and the numerical attributes must be discretized. Illustrate this using screen shots.

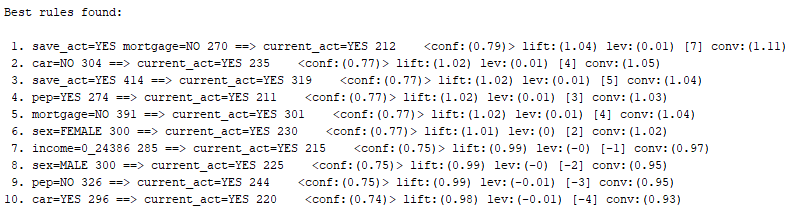


1. Next perform association rule discovery on the transformed data. In WEKA Apriori algorithm interface set "outputItemsets" to "True" so that you can also view the frequent items sets of different sizes in addition to the rules. Illustrate this using a screen shot.



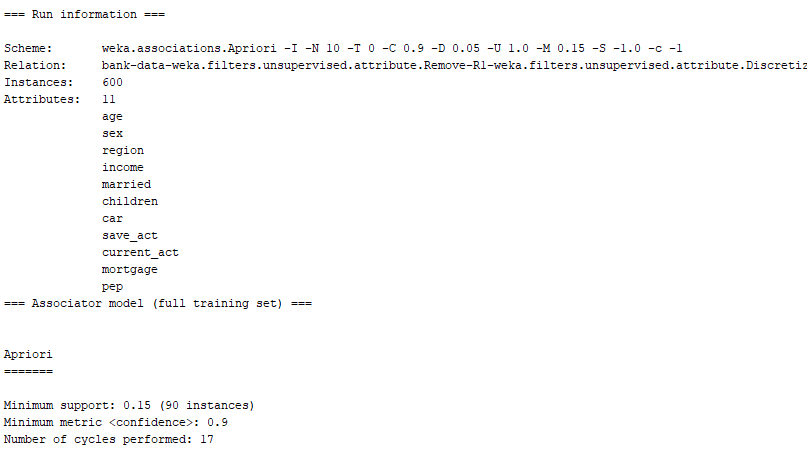


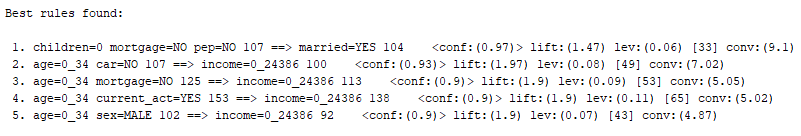




1. Experiment with different parameters so that you get at least 20-30 strong rules (e.g., rules with high confidence which at the same time have relatively good support). Illustrate this using a screen shot.

With 15% support and 90% confidence





With 20% support and 80% confidence

