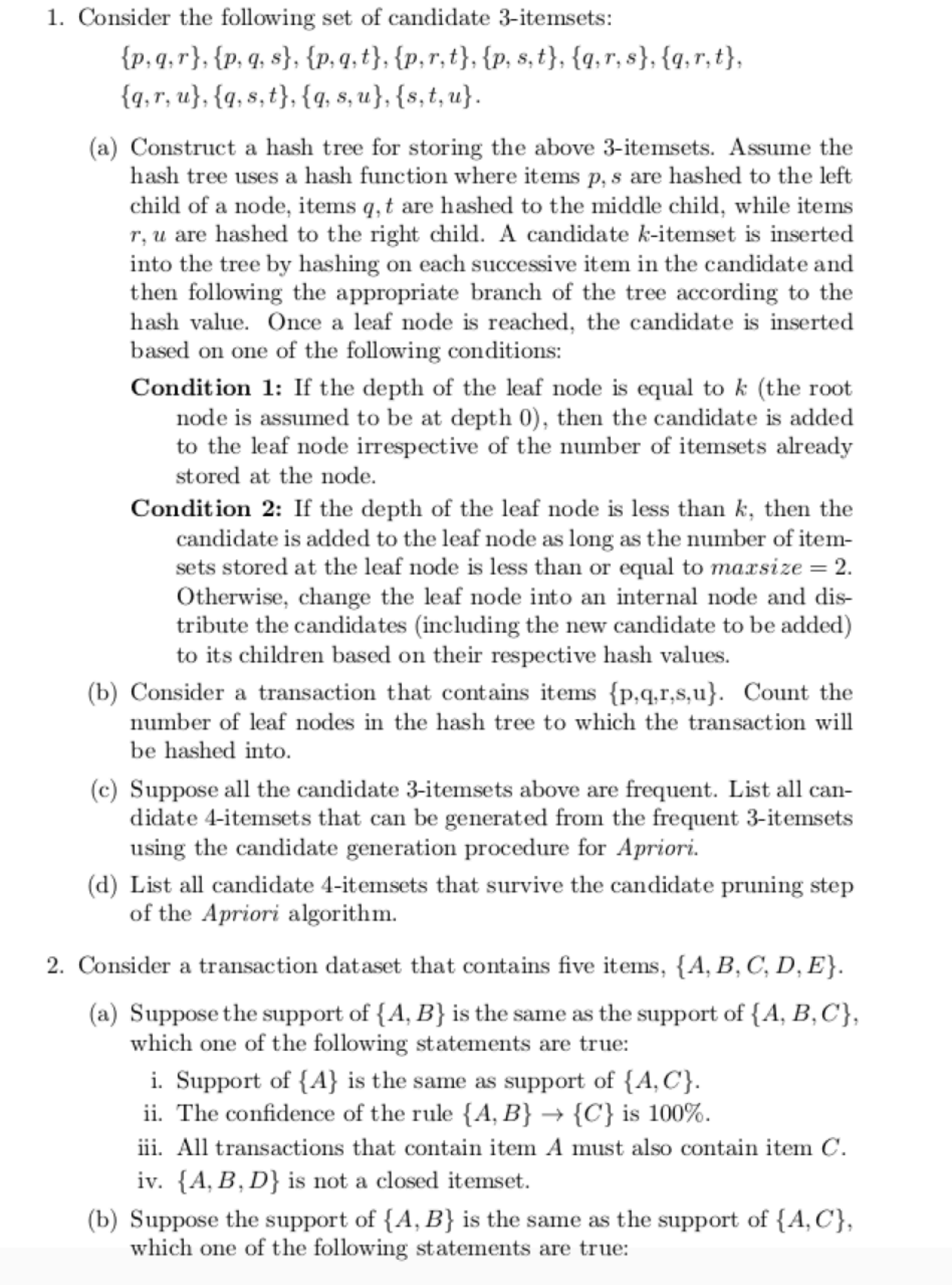
CSE 881: Data Mining (Spring 2020) Homework 3:
Due date: April 30, 2020



1. Consider the following set of candidate 3-itemsets:

   $\{p,q,r\}, \{p,q,s\}, \{p,q,t\}, \{p,r,t\}, \{p,s,t\}, \{q,r,s\}, \{q,r,t\},$
   $\{q,r,u\}, \{q,s,t\}, \{q,s,u\}, \{s,t,u\}.$

   (a) Construct a hash tree for storing the above 3-itemsets. Assume the hash tree uses a hash function where items p, s are hashed to the left child of a node, items q, t are hashed to the middle child, while items r, u are hashed to the right child. A candidate k-itemset is inserted into the tree by hashing on each successive item in the candidate and then following the appropriate branch of the tree according to the hash value. Once a leaf node is reached, the candidate is inserted based on one of the following conditions:

   Condition 1: If the depth of the leaf node is equal to k (the root node is assumed to be at depth 0), then the candidate is added to the leaf node irrespective of the number of itemsets already stored at the node.

   Condition 2: If the depth of the leaf node is less than k, then the candidate is added to the leaf node as long as the number of itemsets stored at the leaf node is less than or equal to $maxsize = 2$. Otherwise, change the leaf node into an internal node and distribute the candidates (including the new candidate to be added) to its children based on their respective hash values.

   (b) Consider a transaction that contains items {p,q,r,s,u}. Count the number of leaf nodes in the hash tree to which the transaction will be hashed into.

   (c) Suppose all the candidate 3-itemsets above are frequent. List all candidate 4-itemsets that can be generated from the frequent 3-itemsets using the candidate generation procedure for *Apriori*.

   (d) List all candidate 4-itemsets that survive the candidate pruning step of the *Apriori* algorithm.

2. Consider a transaction dataset that contains five items, $\{A, B, C, D, E\}$.

   (a) Suppose the support of $\{A, B\}$ is the same as the support of $\{A, B, C\}$, which one of the following statements are true:

   i. Support of $\{A\}$ is the same as support of $\{A, C\}$.
   ii. The confidence of the rule $\{A, B\} \rightarrow \{C\}$ is 100%.
   iii. All transactions that contain item A must also contain item C.
   iv. $\{A, B, D\}$ is not a closed itemset.

   (b) Suppose the support of $\{A, B\}$ is the same as the support of $\{A, C\}$, which one of the following statements are true:

i. All transactions that contain item B must contain item C.

ii. The confidence of the rule $\{A, B\} \rightarrow \{C\}$ is 100%.

iii. The support of $\{A, B\}$ is the same as the support of $\{A, B, C\}$.

iv. $\{A, B, D\}$ is not a closed itemset.

(c) Suppose all the transactions that contain $\{A, B\}$ also contain $\{B, C\}$, which one of the following statements are true:

i. The confidence of the rule $\{B, C\} \rightarrow \{A\}$ is 100%.

ii. The support of $\{A\}$ is the same as the support of $\{C\}$.

iii. $\{A, B, D\}$ is not a closed itemset.

(d) Suppose the confidence of the rules $\{A, B\} \rightarrow C$ and $\{A, B\} \rightarrow D$ are identical, which one of the following statements are true:

i. The confidence of the $\{A, B\} \rightarrow \{C, D\}$ is the same as the confidence of $\{A, B\} \rightarrow \{C\}$.

ii. All transactions that contain {A,B,C} also contain {A,B,D}.

iii. $\{A, B\}$ is not a closed itemset.

(e) Consider the lattice structure shown in Figure 1.

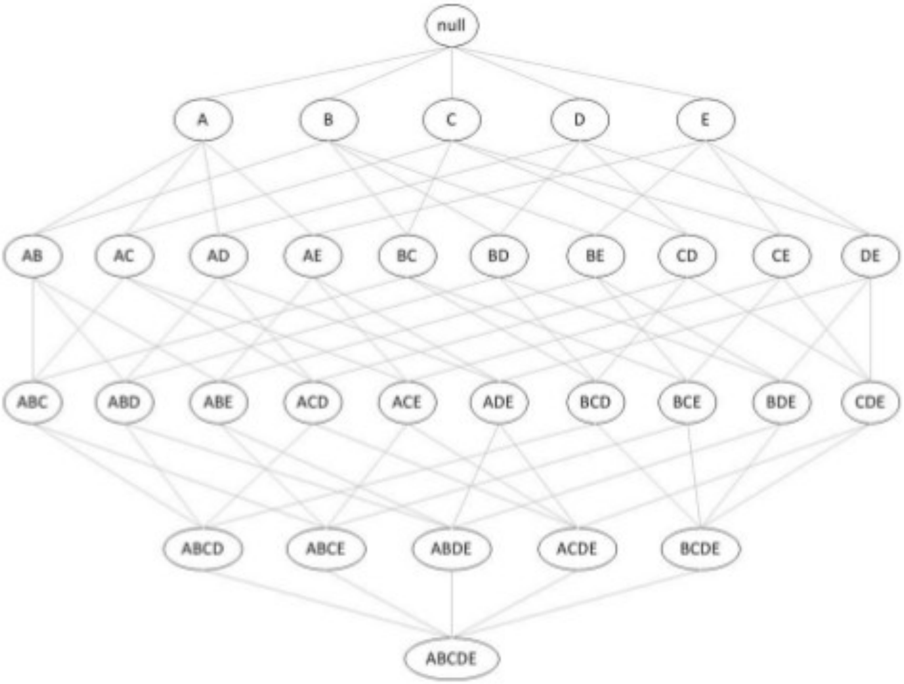


Figure 1: Search space for transaction database that contains 6 items.

For each of the following scenarios, list all the itemsets that are guaranteed to be not closed when:

i. Support of $\{A, B\}$ is equal to support of $\{A, B, C\}$.

ii. All the transactions that contain $\{A, B\}$ is a subset of the transactions that contain $\{C\}$.

3. Download the Apriori software from Christian Borgelt's Web page at http://www.borgelt.net/apriori.html and read carefully the documentation on how to use the software. If you're running the code from one of the CS machines, you can download the 64-bit source code apriori.tar.gz from the Website. After unzipping the file, go to the src subdirectory and type make to create the binary executable files. The command for executing the Apriori code is

   ```
   apriori [options] inputfile outputfile
   ```

   where $[\cdot]$ are the optional command-line arguments.

   (a) Download the votes.tab data file from the class web site. The dataset contains information about the congressional voting records of U.S. House of Representatives. The original data is available from the UCI machine learning repository (https://archive.ics.uci.edu/ml/datasets/Congressional+Voting+Records). The data has been preprocessed to create a transaction file, where each row (line) in the file corresponds to a transaction. Please read the UCI web page to learn more details about the meaning of the items (attributes) created from the original data.

   (b) Apply the Apriori algorithm to create frequent 1-itemsets using a minimum support threshold of 30%. Note you should restrict the output to produce only 1-itemsets (you must exclude 2-itemsets or higher).

   - Specify the command you had used to generate the frequent 1-itemsets.
   - How many frequent itemsets did you find?
   - What are the support for {republican} and {democrat}? Which political party has more representatives (transactions) in the data?
   - Attach the list of frequent itemsets generated with your homework solution (do not submit it as a separate file).

   (c) Create an item appearance file named votes.app. The file can be used to restrict items that can appear as antecedent (body/in) or consequent (head/out) of a rule. Apply the following restrictions:

   - republican and democrat can only appear as the consequent of the rule.
   - Other attributes can only appear as the antecedent of the rule.

   Attach the item appearance file with your homework solution (do not submit it as a separate file).

   (d) Apply the Apriori algorithm to generate association rules from the data using the appearance file you have created in the previous question. Use a minimum support threshold of 30% and minimum confidence of 70%. Also restrict the maximum number of items per rule to be 2. Note that the association rule $X \rightarrow Y$ is displayed as $Y \leftarrow X$ in the output file.

- Specify the command you used to run Apriori to generate the association rules.
- How many rules are generated?
- Based on the rules generated, state which party (republican or democrat) whose representatives will most likely vote yes on the following bills:
  - education-spending
  - adoption-of-the-budget-resolution
  - physician-fee-freeze
  - aid-to-nicaraguan-contras
  - mx-missile
  - el-salvador-aid
- Attach the list of association rules generated with your homework solution (do not submit it as a separate file).

(e) Repeat the previous step by lowering the minimum confidence threshold from 70% to 45% (keeping the rest of the parameters the same as before).

- How many rules are generated?
- Based on the rules generated, state which of the bills where the representatives do not vote along their party lines. You can assume these are the rules that have a confidence between 45% to 55%.