Truc Huynh

Purdue University Fort Wayne

Homework 2

PART I.

Reading Read Fournier-Viger et al., “A survey of itemset mining”, WIREs Data Mining and Knowledge Discovery, Vol. 7:e1207, 2017. Write a (detail) summary of each section following:

1. Variations of the itemset mining problem:

The popular problem includes below

1. A huge amount of item sets (based on minimum support): Too many patterns make data analyzing difficult. Algorithms are designed to extract concise representations of frequent item-sets to solve this matter. A set of a frequent itemset is called “A concise representation” when it is summarizing all frequent itemset in that data frame. Set of frequent item-sets:
   * Closed itemset: frequent item-sets that supersets have the same support.
   * Maximal Item-sets: frequent item-sets which supersets are not frequent.
   * Generator item-sets: frequent item-sets which subsets do not have the same support (also called key item-sets).
2. Introducing constraints to filter less-interesting patterns. There are several constraints, the most popular are:
   * Monotone.
   * The Anti-monotone constraint is one of the easiest and most beneficial to integrate into a Frequent Itemset Mining (FIM) mining algorithm.
   * The succinct constraint can be checked by only looking at the single item that it contains.
   * Convertible constraint: neither monotone nor anti-monotone. Besides, if some strategies are applied by the FIM algorithm, it can be converted into antimonotone constraints.
3. The rare items problem: not every item in real life is equal. However, FIM assumes that all items are equal.
4. Designed to be applied as batch algorithms: In dynamic environments (where changes happen instantly), updated the transaction database(TD) will lead to applying the FIM algorithm again to update the pattern. Some solution has been implemented, the most popular is designed algorithm that can update result in real-time. Some algorithms are:
   * Incremental mining algorithms: when new transactions are inserted, deleted, or modified in a TD, frequent itemset pattern will be automatically updated.
   * Stream mining algorithm: is optimized to process transactions as quickly as possible by calculating the approximate set of a frequent itemset.
   * Interactive mining algorithms: the idea is focusing on the needed item-sets only instead of mining and updating a large number of itemsets that may not all be useful.
5. Database Format:

* Weighted itemset mining is used to find itemsets that have a minimum weight -> mine infrequent weighted itemsets.
* High-utility itemset mining(HUIM): is an extension of weighted itemset mining where both weights and purchase quantities in transactions are considered.
* Uncertain itemset mining: designed to considerable uncertainty about the data
* Fuzzy itemset mining: a well-studied extension of itemset mining.

1. Traditionally applied to find itemsets in a single set of transactions is also a limit of FIM. However, it is often useful to discover the difference between two or more sets of transactions in real-life applications.
2. Other pattern mining problems related to itemset:

This session shows some other important pattern mining problems and how to solve them:

* 1. An association rule mining algorithm works best with a minimum support threshold mins up and a minimum confidence threshold min conf → discover patterns representing strong associations between items → Limitation: May find frequent itemsets that are weakly correlated.
  2. Sequential pattern mining: include discovering sequences frequently appearing in a set of sequences → Limitation: Similar to FIM’s problem, and to discover sequential rules → Application: more effective than sequential patterns for some tasks involving prediction → a frequent partial order can summarize several sequential patterns
  3. Episode mining: → Limitation: mine in a single sequence rather than in a set of sequences → used to analyze web-click streams, telecommunication data, sensor readings, sequences of events on an assembly line, and network traffic data
  4. Periodic pattern mining is an algorithm for discovering periodic frequent patterns → Application: discovering periodic behavior of customers, and finding recurring events (stock market analysis, market analysis, and bioinformatics).
  5. Subgraph mining: goal is to discover frequent subgraphs in a database of graphs rather than frequent item-sets in transactions → Limitation: Similar to FIM’s problem, search space is generally very large → Application: mine closed and maximal frequent subgraphs.

PART II:

**Question 1:**

(1) All non-empty subsets of a frequent itemset must also be frequent:

Proof:

Let 𝑆 ⊆ 𝐼 be a frequent itemset, i.e. 𝑠upport (𝑆) ≥ minSup

Let ∅ ≠ 𝑆′ ⊆ S

Then support(𝑆′) ≥ 𝑏) support(S)

≥

i.e. S’ is a frequent itemset

(Seidl., n.d.)

For example:

If {beer, diaper, nuts} is frequent, so is {beer, diaper}. i.e., every transaction having {beer, diaper, nuts} also contains {beer, diaper} (P. Tan et al., n.d.)

(2) Prove that the support of any nonempty subset X′ of itemset X must be at least as great as the support of X. For the proof, give a brief explanation with an example data.

Proof:

Let ∅ ≠ 𝑆′ ⊆ 𝑆 ⊆ 𝐼

For any transaction 𝑇 ⊆ 𝐼 in database 𝐷, we have:

𝑆 ⊆ 𝑇 ⇒ 𝑆’ ⊆ T

Thus, it holds that

{𝑇 ∈ 𝐷 | 𝑆 ⊆ 𝑇} ⊆ {𝑇 ∈ 𝐷 | 𝑆’ ⊆ T}

And consequently:

Support(S)= {𝑇 ∈ 𝐷 | 𝑆 ⊆ 𝑇} ≤ {𝑇 ∈ 𝐷 | 𝑆’ ⊆ T} = support(S’)

For example:

**Question 2:**

1. Which data set(s) will produce the most number of frequent itemsets? Explain your answer.

Answer: Data set (e) because it generate the most number of frequent itemsets over the other data sets.

1. Which data set(s) will produce the fewest number of frequent itemsets? Explain your answer.

Answer: Data set (d) because it does not contain any frequent itemset in the diagram.

1. Which data set(s) will produce the longest frequent itemset? Explain your answer.

Answer: Dataset (e) because it generate the longest of frequent itemsets over the other data sets.

1. Which data set(s) will produce frequent itemsets with highest maximum support? Explain your answer.

Answer: Data set (b) because it contain the itemsets with highest maximum support

1. Which data set(s) will produce frequent itemsets containing items with wide-varying support levels (i.e., items with mixed support, ranging from less than 0.2 to more than 0.7). Explain your answer.

Answer: Data set (e) because it generate the most number of frequent itemsets with wide-varying support levels over the other data sets.

**Question 3:** Apply min\_support threshold is 0.6 (min\_support\_count=3)

Database D

Scan D

|  |  |
| --- | --- |
| **TID** | **Items** |
| **1**  **2**  **3**  **4**  **5** | {M, O, N, K, E, Y}  {D, O, N, K, E, Y}  {M, A, K, E}  {U, C, K, Y}  {C, O, K, I, E} |

C1

C1

L1

Scan D

|  |  |
| --- | --- |
| **Itemsets** | **Sup\_count** |
| {K}  {E}  {O}  {Y} | 100%  80%  60%  60% |

|  |  |
| --- | --- |
| **Itemsets** | **Sup\_count** |
| ~~{A}~~  ~~{C}~~  ~~{D}~~  {E}  {K}  ~~{M}~~  ~~{N}~~  {O}  ~~{U}~~  {Y} | ~~20%~~  ~~40%~~  ~~20%~~  80%  100%  ~~40%~~  ~~40%~~  60%  ~~20%~~  60% |

|  |
| --- |
| **Itemsets** |
| {A}  {C}  {D}  {E}  {K}  {M}  {N}  {O}  {U}  {Y} |

C2

C2

C2

|  |
| --- |
| **Itemsets** |
| {KE}  {KO}  {KY}  {EO}  {EY}  {OY} |

|  |  |
| --- | --- |
| **Itemsets** | **Sup\_count** |
| {KE}  {KO}  {KY}  {EO}  ~~{EY}~~  ~~{OY}~~ | 80%  60%  60%  60%  ~~40%~~  ~~40%~~ |

|  |
| --- |
| **Itemsets** |
| {KE}  {KO}  {KY}  {EO}  {EY}  {OY} |

PRUNE C2

Scan D

L2

|  |  |
| --- | --- |
| **Itemsets** | **Sup\_count** |
| {KE}  {KO}  {KY}  {EO} | 80%  60%  60%  60% |

C3

PRUNE C3

C3

C3

|  |
| --- |
| **Itemsets** |
| {KEO}  {KEY}  {KOY}  {EOY}} |

|  |  |
| --- | --- |
| **Itemsets** | **Sup\_count** |
| {KEO}  ~~{KEY}~~  ~~{KOY}~~  ~~{EOY}~~ | 60%  ~~40%~~  ~~40%~~  ~~40%~~ |

|  |
| --- |
| **Itemsets** |
| {KEO}  {KEY}  {KOY}  {EOY} |

Scan D

L3

|  |  |
| --- | --- |
| **Itemsets** | **Sup\_count** |
| {KEO} | 60% |

C4 is Empty

Question 4:

Diagram

Description automatically generated

**Question 5:**

Show the procedure to find all frequent itemsets from the above transaction dataset using FP-growth (with FP-tree) algorithm. Assume the min\_support threshold is 0.6 (min\_support\_count=3). Show (1) Flist, (2) the transaction data with ordered frequent items, (3) FP-tree, (4) Conditional pattern bases, (5) Conditional FP-tree per each pattern base, and (6) the frequent itemsets generated from each conditional FP-tree.

|  |
| --- |
| **(Ordered) frequent items** |
| {K, E, O, Y}  {K, E, O, Y}  {K, E}  {K, Y}  {K, E, O} |

|  |  |
| --- | --- |
| **TID** | **Items** |
| **1**  **2**  **3**  **4**  **5** | {M, O, N, K, E, Y}  {D, O, N, K, E, Y}  {M, A, K, E}  {U, C, K, Y}  {C, O, K, I, E} |

|  |  |
| --- | --- |
| **Header Table** | |
| **Item** | **frequency** |
| K  E  O  Y  ~~C~~  ~~M~~  ~~N~~  ~~A~~  ~~D~~  ~~U~~ | 5  4  3  3  ~~2~~  ~~2~~  ~~2~~  ~~1~~  ~~1~~  ~~1~~ |

Initial FP tree

{}

|  |  |  |
| --- | --- | --- |
| **Conditional Pattern Base** | | |
| **Item** | **Cond.pattern base** | **Conditional**  **FP-Tree** | **Frequent Pattern Generated** |
| Y | {K, E, O:2},  {K:1} | <K:2, ~~E:2, O:2~~>  <K:1> | {K, Y: 3} |
| O | {K, E:3} | <K:3, E:3> | {K, O: 3}, {E, O: 3}, {K, E, O: 3} |
| E | {K:4} | <K:4> | {K, E: 4} |

1. Scan DB, find frequent itemset: L1={E,K,O,Y}
2. Sort frequent item in frequent descending order, F-list. F-list=K-E-O-Y

|  |  |
| --- | --- |
| **Header Table** | |
| **Item** | **Cond.pattern base** |
| K  E  O  Y | {}  {K}:4  {K, E}:3  {K, E, O}:2, {K}:1 |

Y:1

Y:2

K:5

E:4

O:3

**Question 6:**

Find (1) all the maximal frequent itemsets and (2) all the closed frequent itemsets when the min\_support threshold is 0.6 (min\_support\_count=3).

Closed: {K}, {E}, {KE}, {K, E, O}

|  |  |
| --- | --- |
| **Frequent Itemsets** | **Support** |
| {K}  {E}  {O}  {Y}  {K, E}  {K, O}  {K, Y}  {E, O}  {K, E, O} | 1  0.8  0.6  0.6  0.8  0.6  0.6  0.6  0.6 |

Maximal: {K, E, O}

Question 7:when the confidence threshold is 0.7

|  |  |
| --- | --- |
| **Frequent Itemsets** | **Support** |
| {K, E, O} | 0.6 |

c ({K, E} → {O}) =σ{K, E, O}/ σ{K, E} = 3/4= 0.75

c ({K, O} → {E}) =σ{K, E, O}/ σ{K, O} = 3/3= 1

c ({E, O} → {K}) =σ{K, E, O}/ σ{E, O} = 3/3= 1

~~c ({K} → {E,O}) =σ{K, E, O}/σ{K} = 3/5 = 0.6~~ (< confidence threshold)

c ({E} → {K,O}) =σ{K, E, O}/ σ{E}= 3/4= 0.75

c ({O}→ {K,E}) =σ{K, E, O}/ σ{O}= 3/3=1

Strong rules:

|  |  |
| --- | --- |
| **Frequent Itemsets** | **CONFIDENCE** |
| K, E → O  K, O → E  E, O → K | 0.75  1  1 |
| ~~K → E, O~~  E → K, O  O → K, E | ~~0.6~~  0.75  1 |

**Question 8:**

|  |  |
| --- | --- |
| **TID** | **Items** |
| **1**  **2**  **3**  **4**  **5** | {M, O, N, K, E, Y}  {D, O, N, K, E, Y}  {M, A, K, E}  {U, C, K, Y}  {C, O, K, I, E} |

Contingency table K → Y:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Y |  | sum |
| K | 3 |  |  |
|  | 3 |  |  |
| sum |  |  | N= 13 |

Contingency table Y → K:

|  |  |  |  |
| --- | --- | --- | --- |
|  | K |  | sum |
| Y | 3 |  |  |
|  |  |  |  |
| sum |  |  | N= 13 |

|  |  |  |
| --- | --- | --- |
|  | **K → Y** | **Y → K** |
| **Support** s(K, Y) | s(K ,Y) = = 3/13= 0.23 | s(Y, K) = = = 3/13= 0.23 |
| **Confidence** c (K → Y) | c (K → Y) = = 3/8=0.375 | c (Y → K) ==3/6=0.5 |
| **Lift** | Lift(K,Y)= s(K ,Y)/(s(K)\*s(Y))  = = 3\*13/(8\*6)=0.8125 < 1:  Negative associated | Lift(Y,K)= s(Y ,K)/(s(Y)\*s(K))  = = 3\*13/(6\*8)=0.8125 < 1:  Negative associated |
| **Leverage** | Leverage (K ,Y)= s(K ,Y) - (s(K)\*s(Y))=0.23-(8/13\*6/13)=0.23-0.28=-0.05 < 0 | Leverage (Y ,K)= s(Y ,K) -(s(Y)\*s(K))= 0.23-(6/13\*8/13) =0.23-0.28=-0.05 < 0 |
| **Conviction** | *max* (0.862, 0.769)  = ½ (0.862+ 0.769+|0.862- 0.769|)  = 0.862 | *max* (0.769,0.862)  = 0.862 |

**Question 9:**

Present the transaction data with the vertical data layout:

|  |  |
| --- | --- |
| **TID** | **Items** |
| **1**  **2**  **3**  **4**  **5** | {M, O, N, K, E, Y}  {D, O, N, K, E, Y}  {M, A, K, E}  {U, C, K, Y}  {C, O, K, I, E} |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Vertical Data Layout** | | | | | | | | | |
| **K** | **E** | **O** | **Y** | **C** | **M** | **N** | **A** | **D** | **U** |
| 1  2  3  4  5 | 1  2  3  5 | 1  2  5 | 1  2  4 | 4  5 | 1  3 | 1  2 | 3 | 2 | 4 |

References:

P. Tan et al., (n.d.) “Introduction to Data Mining”, Chapter 2. Retrieved from https://purdue.brightspace.com/d2l/le/content/127433/viewContent/3920610/View

Seidl., T. (n.d.) Exercise 3: Frequent Itemset Mining, Knowledge Discovery in Databases I SS 2016. Retrieved from: https://www.dbs.ifi.lmu.de/Lehre/KDD/SS16/uebung/blatt03\_sol.pdf