Chapter 4

Association Rule

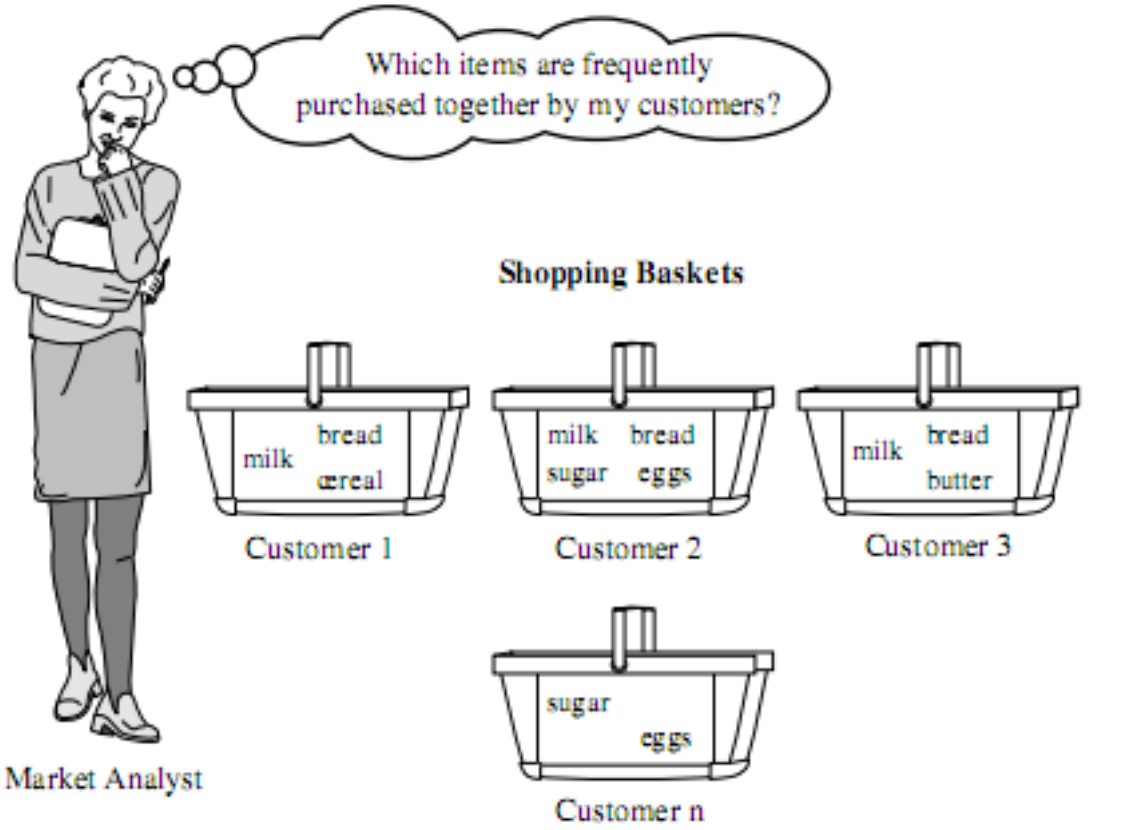
Association Rule(เหมืองข้อมูลแบบกฎความสัมพันธ์)

- เป็นเทคนิคหนึ่งของ Data Mining คือการค้นหาความสัมพันธ์ของข้อมูลจากข้อมูลขนาดใหญ่ที่มีอยู่เพื่อนำไปหารูปแบบที่เกิดขึ้นบ่อยๆ (frequent pattern) และใช้ในการวิเคราะห์ความสัมพันธ์หรือทำนายปรากฏการณ์ต่างๆ
- ฐานข้อมูลที่ใช้ในการทำเหมืองความสัมพันธ์ (Association Mining) มักเป็นฐานข้อมูลประเภท Transaction Database
- ผลลัพธ์ที่ได้เป็นกฎความสัมพันธ์ (Association Rule) สามารถเขียนได้ในรูปเซตของรายการที่เป็นเหตุ ไปสู่เซตของรายการที่เป็นผล ซึ่งมีรากฐานมาจากการวิเคราะห์ตะกร้าตลาด (Market Basket Analysis) เช่น ลูกค้าที่ซื้อผ้าอ้อมส่วนใหญ่จะซื้อเบียร์ด้วย
- ข้อมูลที่นำมาใช้จะอยู่ในรูปแบบ Nominal หรือ Ordinal เท่านั้น

การนำเทคนิคไปประยุกต์ใช้กับงานจริง

- ระบบแนะนำหนังสือให้กับลูกค้าแบบอัตโนมัติของ Amazon หมายถึงข้อมูลการสั่งซื้อทั้งหมดจะถูกนำมาประมวลผลเพื่อหาความสัมพันธ์ของข้อมูล เช่น ลูกค้าที่ซื้อหนังสือเล่มหนึ่งๆ มักจะซื้อหนังสือเล่มใดพร้อมกันด้วยเสมอ ความสัมพันธ์ที่ได้จากกระบวนการนี้สามารถนำไปใช้คาดเดาได้ว่าควรแนะนำหนังสือเล่มใดเพิ่มเติมให้กับลูกค้า

Association Rule



Market basket analysis.

Association Rule

- การวิเคราะห์ตะกร้าตลาด เป็นรูปแบบที่ใช้เพื่อหากลุ่มสิ่งของที่น่าจะปรากฏร่วมกันใน transaction หนึ่งๆ ซึ่งมักเป็น transaction ณ จุดขาย ผลลัพธ์ที่ได้สามารถแสดงได้ด้วยกฎ ซึ่งบอกความเป็นไปได้ของการซื้อสินค้าต่างๆร่วมกัน
- การวิเคราะห์ตะกร้าตลาด มีบทบาทสำคัญต่ออุตสาหกรรมการค้าปลีก (retail industry) ซึ่งใช้สารสนเทศ ศึกษาพฤติกรรมของลูกค้า
  - จัดพื้นที่ร้านค้า
  - จัดวางสินค้าร่วมกันเพื่อส่งเสริมการขาย
  - การวางแผนการส่งเสริมการขายและตั้งราคาผลิตภัณฑ์

Support & Confidence

- Support คือ ตัววัดประสิทธิภาพสำหรับสินค้า (Item)
  สมมติการซื้อสินค้าครั้งหนึ่งมี จำนวน Transaction = 3 และมีรายการดังนี้

Product	Apple	Beer	Cereal	Diapers	Eggs
Apple	2				
Beer	2	3			
Cereal	1	1	1		
Diapers	2	2	1		
Eggs	1	2	0	1	2

support

support

ค่า support (“Beer”) = 3/3 = 1

ค่า support (“Apple” AND “Diaper”) = 2/3 = 0.67

Support & Confidence (ต่อ)

- Confidence คือ ตัววัดประสิทธิภาพของกฎความสัมพันธ์ (Association rule) เช่น
  - R1 : Buy(“Apple”) => Buy(“Diapers”)
  - Confidence (R1) = Support(“Apple”,“Diapers”) / Support(“Apple”) = 2/2 = 1 (100%)
  - ทุกครั้งที่มีการซื้อ Apple ลูกค้าก็จะซื้อ Diapers ไปด้วยทุกๆ ครั้ง

ID	Product
1	Apple, Beer, Cereal, Diapers
2	Apple, Beer, Diapers, Eggs
3	Beer, Eggs

Definition: Frequent Itemset

- Itemset
  - A collection of one or more items
  - Example: {Milk, Bread, Diaper}
- k-itemset
  - An itemset that contains k items
- Support count (**σ**)
  - Frequency of occurrence of an itemset
  - E.g. **σ**({Milk, Bread, Diaper}) = 2
- Support (ตัววัดประสิทธิภาพสำหรับสินค้า (item))
  - Fraction of transactions that contain an itemset
  - E.g. s({Milk, Bread, Diaper}) = 2/5
- Frequent Itemset
  - An itemset whose support is greater than or equal to a minsup threshold

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Definition: Association Rule

- Association Rule
  - An implication expression of the form X ⟶ Y, where X and Y are itemsets
  - Example: {Milk, Diaper} ⟶ {Beer}
- Rule Evaluation Metrics
  - Support (s)
    - Fraction of transactions that contain both X and Y
  - Confidence (c) (ตัววัดประสิทธิภาพของกฎความสัมพันธ์)
    - Measures how often items in Y appear in transactions that contain X

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Example:

$$\{\text{Milk}, \text{Diaper}\} \Rightarrow \text{Beer}$$

$$s = \frac{\sigma(\text{Milk}, \text{Diaper}, \text{Beer})}{|\text{T}|} = \frac{2}{5} = 0.4$$

$$c = \frac{\sigma(\text{Milk}, \text{Diaper}, \text{Beer})}{\sigma(\text{Milk}, \text{Diaper})} = \frac{2}{3} = 0.67$$

Example

- ถ้าเราได้กฎความสัมพันธ์จากฐานข้อมูลการซื้อสินค้า เป็น

"{B, C} => {A} (ค่า support = 50%, ค่า confidence = 80%)" หมายความว่า จะมีการซื้อ A, B และ C พร้อมกัน 50 ทรานแซคชั่น และ 80 เปอร์เซ็นต์ของลูกค้าที่ซื้อ B พร้อมกับ C จะซื้อ A ไปด้วย

Association Rule Mining Task

- Given a set of transactions T, the goal of association rule mining is to find all rules having
  - support ≥ minsup threshold
  - confidence ≥ minconf threshold
- Brute-force approach:
  - List all possible association rules
  - Compute the support and confidence for each rule
  - Prune rules that fail the minsup and minconf thresholds

Example

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Example of Rules:

$\{Milk,Diaper\} \rightarrow \{Beer\}$ (s=0.4, c=0.67)

$\{Milk,Beer\} \rightarrow \{Diaper\}$ (s=0.4, c=1.0)

$\{Diaper,Beer\} \rightarrow \{Milk\}$ (s=0.4, c=0.67)

$\{Beer\} \rightarrow \{Milk,Diaper\}$ (s=0.4, c=0.67)

$\{Diaper\} \rightarrow \{Milk,Beer\}$ (s=0.4, c=0.5)

$\{Milk\} \rightarrow \{Diaper,Beer\}$ (s=0.4, c=0.5)

Mining Association Rules

- Two-step approach:

  1. Frequent Itemset Generation

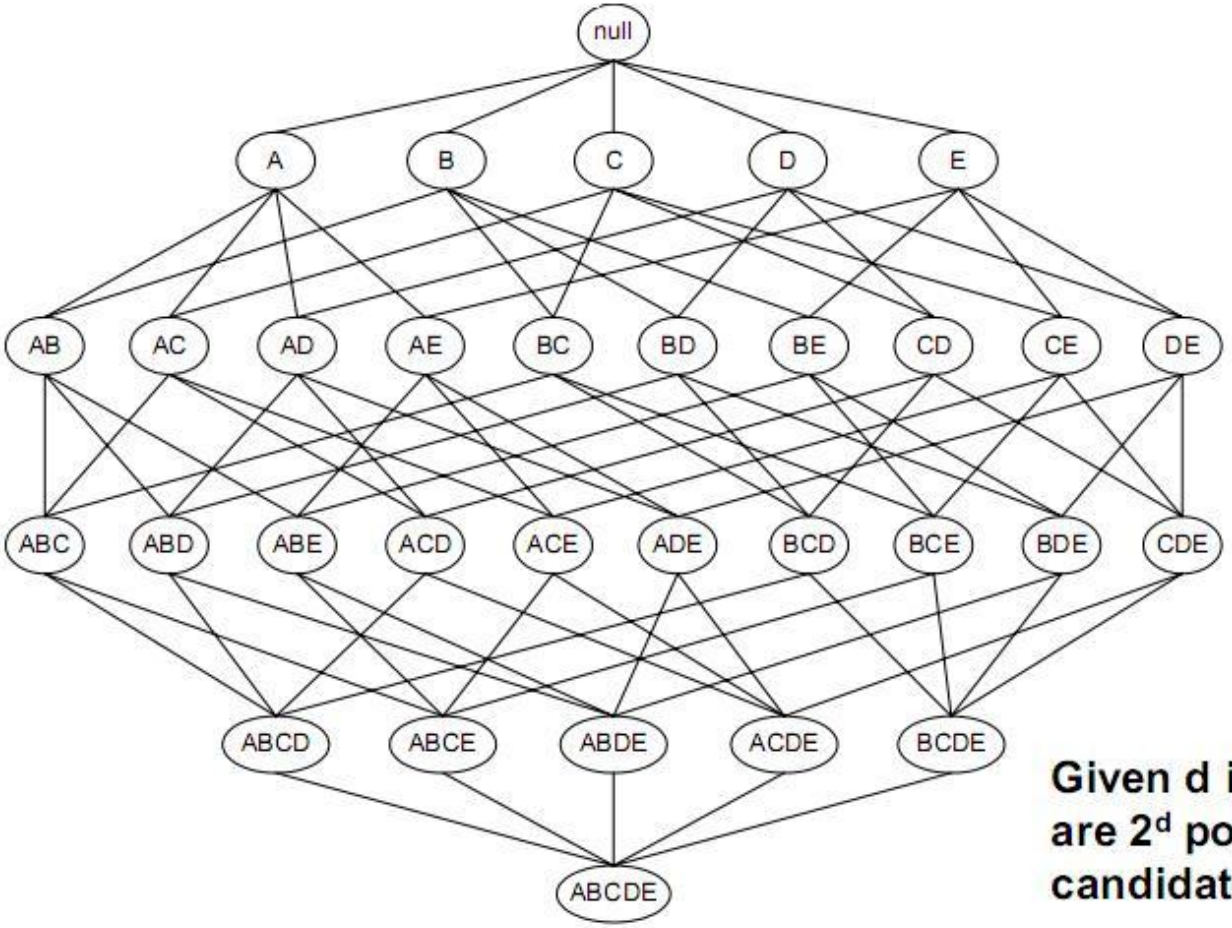
     – Generate all itemsets whose support ≥ minsup

  2. Rule Generation

     – Generate high confidence rules from each frequent itemset,

       where each rule is a binary partitioning of a frequent itemset

Frequent Itemset Generation



Given d items, there are 2^d possible candidate itemsets

The Apriori Algorithm

- Apriori เป็นอัลกอริทึมพื้นฐานที่ใช้ในการหาความสัมพันธ์ของข้อมูล โดยใช้หลักการค้นหาแบบบวงกว้างก่อนนับทรานแซคชัน ซึ่งจะทำการสร้างและตรวจสอบเซตไอเท็มที่เกิดขึ้นบ่อยทีละชั้น โดยเริ่มจากเซ็ตไอเท็มที่มีจำนวนสมาชิกเท่ากับหนึ่งถ้าเซ็ตไอเท็มใดมีค่าสนับสนุนน้อยกว่าค่าสนับสนุนที่กำหนดก็จะตัดเซตไอเท็มนั้นออก ไม่นำไปสร้างเซ็ตไอเท็มในชั้นต่อไป การทำงานของอัลกอริทึมจะวนไปเรื่อยๆ จนกระทั่งไล่ทุกระดับชั้นหรือไม่เหลือเซ็ตไอเท็มในชั้นต่อไป ในการนับจำนวนทรานแซคชันอัลกอลิทึม Apriori จะทำการไล่ทรานแซคชันครั้งเดียวในแต่ละระดับชั้น ในการตรวจดูว่าทรานแซคชันนั้นบรรจุเซตไอเท็มใดบ้าง เพื่อความรวดเร็วจะเก็บเซตไอเท็มในแต่ละระดับชั้นทั้งหมดไว้ในโครงสร้าง Hash Tree จุดเด่นของอัลกอริทึมนี้อยู่ที่ความสามารถในความเร็วของการค้นหาไอเท็มเซตที่ปรากฏบ่อย ด้วยการละเว้นการพิจารณาไอเท็มเซตที่ปรากฏซ้ำด้วยความถี่ที่ต่ำกว่าเกณฑ์

ตัวอย่างการใช้งานอัลกอริทึม Apriori

- มีการกำหนดค่าสนับสนุนขั้นต่ำ (Minimum Support)
- มีการกำหนดค่าความเชื่อมั่นขั้นต่ำ (Minimum Confidence)
- ในการกำหนดขั้นต่ำทั้งสองค่านี้ จะขึ้นอยู่กับผู้ใช้ระบบเป็นผู้กำหนดเอง หรือจะใช้เชี่ยวชาญ (Expert user) เป็นผู้กำหนดให้ก็ได้ โดยกฎความสัมพันธ์ที่ได้นั้นจะต้องมีค่าสนับสนุน (Support) และค่าความเชื่อมั่น (Confidence) ไม่น้อยกว่าค่าขั้นต่ำที่ได้กำหนดเอาไว้ข้างต้น
- ค่าสนับสนุน (Support) คือ เปอร์เซ็นต์ของจำนวน Itemsets ทั้งหมดที่เกิดขึ้นในฐานข้อมูล
- ค่าความเชื่อมั่น (Confidence) คือ เปอร์เซ็นต์ของจำนวน Itemsets ทั้งหมดที่เกิดขึ้นในฐานข้อมูล ต่อ จำนวน Itemsets ที่เกิดขึ้นทางด้านซ้ายมือของกฎ

The Apriori Algorithm: Basics

Key Concepts :

- Frequent Itemsets: The sets of item which has minimum support (denoted by L_i for i^{th} -Itemset).
- Apriori Property: Any subset of frequent itemset must be frequent.
- Join Operation: To find L_k , a set of candidate k-itemsets is generated by joining L_{k-1} with itself.

The Apriori Algorithm

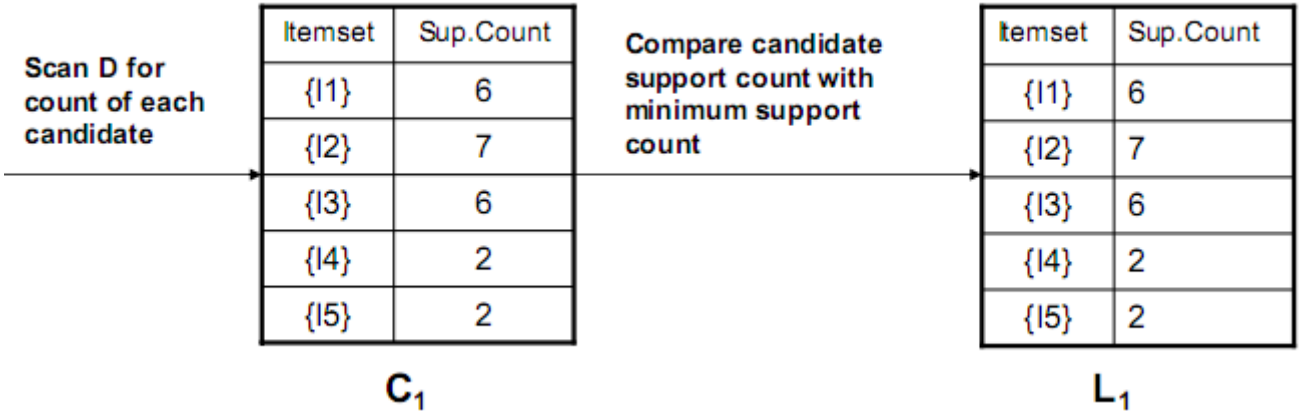
- Find the frequent itemsets: the sets of items that have minimum support
  - A subset of a frequent itemset must also be a frequent itemset i.e., if {AB} is a frequent itemset, both {A} and {B} should be a frequent itemset
  - Iteratively find frequent itemsets with cardinality from 1 to k (k-itemset)
- Use the frequent itemsets to generate association rules

The Apriori Algorithm: Example

TID	List of Items
T100	I1, I2, I5
T100	I2, I4
T100	I2, I3
T100	I1, I2, I4
T100	I1, I3
T100	I2, I3
T100	I1, I3
T100	I1, I2 ,I3, I5
T100	I1, I2, I3

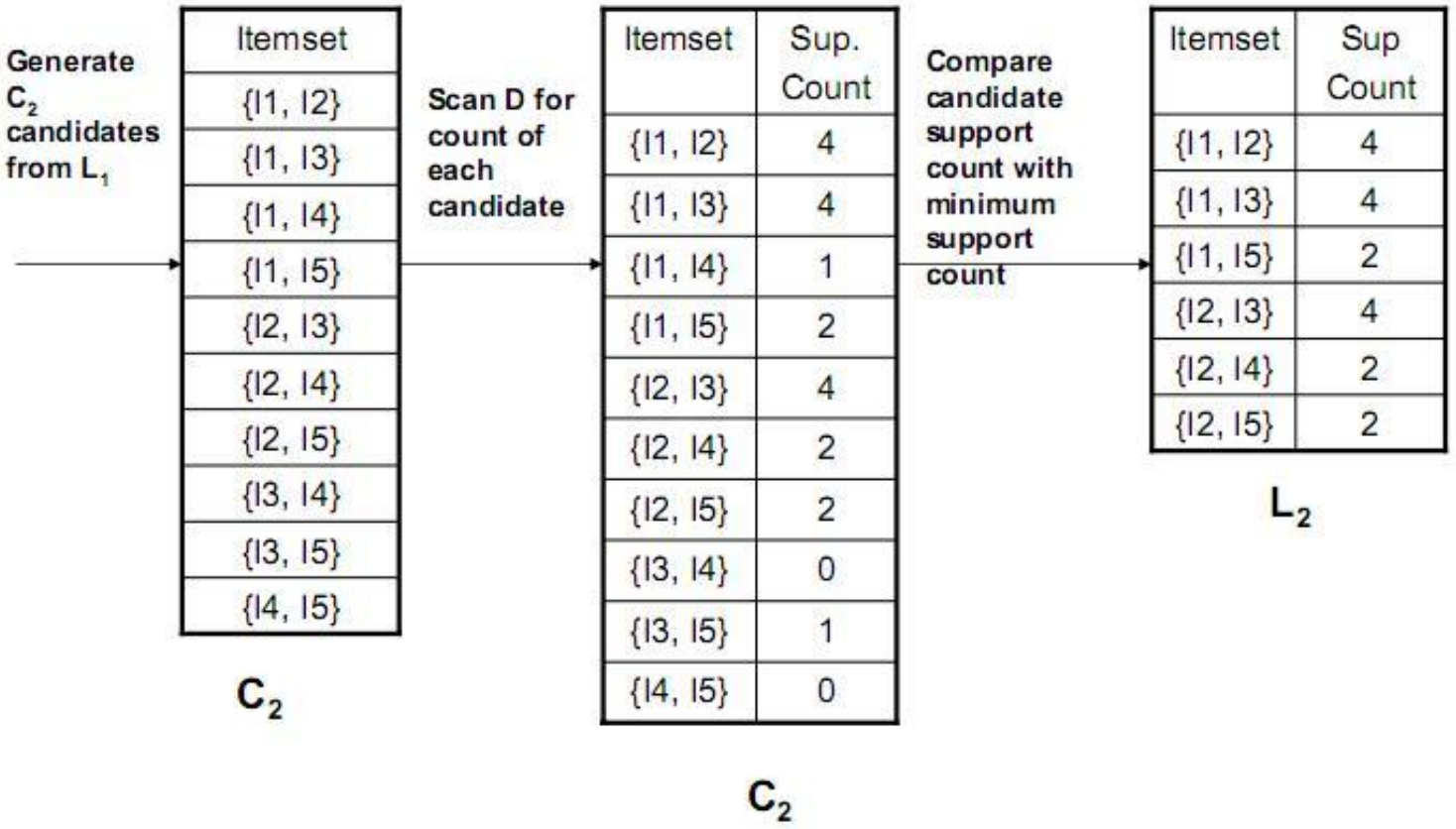
- Consider a database, D , consisting of 9 transactions.
- Suppose min. support count required is 2 (i.e. min_sup = 2/9 = 22 %)
- Let minimum confidence required is 70%.
- We have to first find out the frequent itemset using Apriori algorithm.
- Then, Association rules will be generated using min.support & min.confidence.

Step 1: Generating 1-itemset Frequent Pattern



- The set of frequent 1-itemsets, L1 , consists of the candidate 1-itemsets satisfying minimum support.
- In the first iteration of the algorithm, each item is a member of the set of candidate.

Step 2: Generating 2-itemset Frequent Pattern

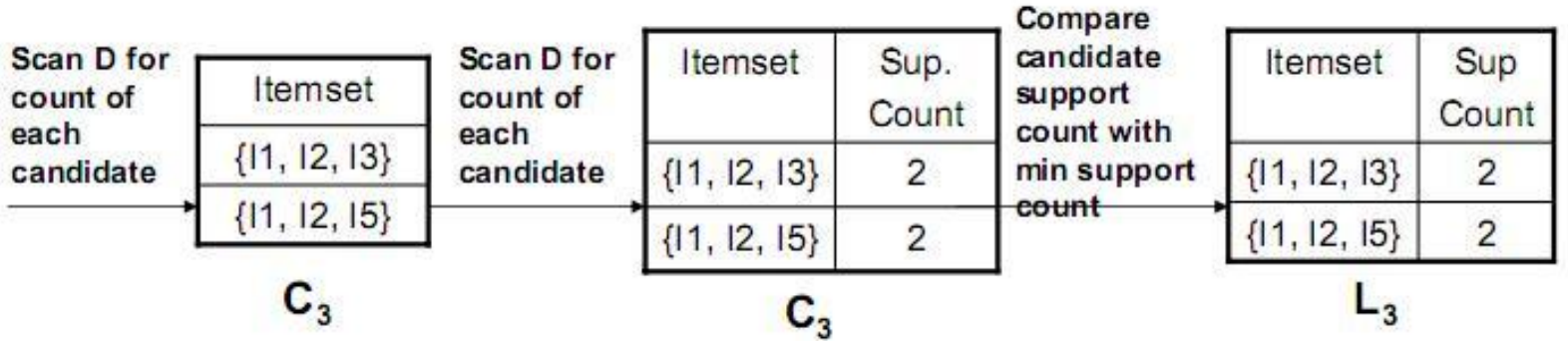


Step 2: Generating 2-itemset Frequent Pattern

- To discover the set of frequent 2-itemsets, L2 , the algorithm uses L1 Join L1 to generate a candidate set of 2-itemsets, C2.
- Next, the transactions in D are scanned and the support count for each

  candidate itemset in C2 is accumulated (as shown in the middle table).
- The set of frequent 2-itemsets, L2 , is then determined, consisting of those candidate 2-itemsets in C2 having minimum support.

Step 3: Generating 3-itemset Frequent Pattern



- The generation of the set of candidate 3-itemsets, C3 ,involves use of the Apriori Property.
- In order to find C3, we compute L2 Join L2.
- C3 = L2 Join L2 = {{I1, I2, I3}, {I1, I2, I5}, {I1, I3, I5}, {I2, I3, I4}, {I2, I3, I5}, {I2, I4, I5}}.
- Now, Join step is complete and Prune step will be used to reduce the size of C3. Prune step helps to avoid heavy computation due to large Ck.

Step 3: Generating 3-itemset Frequent Pattern

- Based on the Apriori property that all subsets of a frequent itemset must also be frequent, we can determine that four latter candidates cannot possibly be frequent. How ?
- For example , lets take {I1, I2, I3}. The 2-item subsets of it are {I1, I2}, {I1, I3} & {I2, I3}. Since all 2-item subsets of {I1, I2, I3} are members of L2, We will keep {I1, I2, I3} in C3.
- Lets take another example of {I2, I3, I5} which shows how the pruning is performed. The 2-item subsets are {I2, I3}, {I2, I5} & {I3,I5}.
- BUT, {I3, I5} is not a member of L2 and hence it is not frequent violating Apriori Property. Thus We will have to remove {I2, I3, I5} from C3.
- Therefore, C3 = {{I1, I2, I3}, {I1, I2, I5}} after checking for all members of result of Join operation for Pruning.
- Now, the transactions in D are scanned in order to determine L3, consisting of those candidates 3-itemsets in C3 having minimum support

Step 4: Generating 4-itemset Frequent Pattern

- The algorithm uses L3 Join L3 to generate a candidate set of 4-itemsets, C4. Although the join results in {{I1, I2,I3, I5}}, this itemset is pruned since its subset {{I2, I3, I5}}is not frequent.
- Thus, C4 = φ , and algorithm terminates, having found all of the frequent items. This completes our Apriori Algorithm.
- What's Next ?

  These frequent itemsets will be used to generate strong association rules (where strong association rules satisfy both minimum support and minimum confidence.)

Step 5: Generating Association Rules from Frequent Itemsets

- Procedure:
- For each frequent itemset "l", generate all nonempty subsets of l.
- For every nonempty subset s of l, output the rule "s -> (l-s)" if support_count(l) / support_count(s) >= min_conf where min_conf is minimum confidence threshold.
- Back To Example:

We had L = {{I1}, {I2}, {I3}, {I4}, {I5}, {I1,I2}, {I1,I3}, {I1,I5}, {I2,I3}, {I2,I4}, {I2,I5}, {I1,I2,I3}, {I1,I2,I5}}.

– Lets take l = {I1,I2,I5}.

– Its all nonempty subsets are {I1,I2}, {I1,I5}, {I2,I5}, {I1}, {I2}, {I5}.

Step 5: Generating Association Rules from Frequent Itemsets

- Let minimum confidence threshold is , say 70%.
- The resulting association rules are shown below, each listed with its confidence.
  - R1: I1 ^ I2 → I5
    - Confidence = sc{I1,I2,I5}/sc{I1,I2} = 2/4 = 50%
    - R1 is Rejected.
  - R2: I1 ^ I5 → I2
    - Confidence = sc{I1,I2,I5}/sc{I1,I5} = 2/2 = 100%
    - R2 is Selected.
  - R3: I2 ^ I5 → I1
    - Confidence = sc{I1,I2,I5}/sc{I2,I5} = 2/2 = 100%
    - R3 is Selected.

Step 5: Generating Association Rules from Frequent Itemsets

– R4: I1 → I2 ^ I5

- Confidence = sc{I1,I2,I5}/sc{I1} = 2/6 = 33%
- R4 is Rejected.

–R5: I2 → I1 ^ I5

- Confidence = sc{I1,I2,I5}/{I2} = 2/7 = 29%
- R5 is Rejected.

– R6: I5 → I1 ^ I2

- Confidence = sc{I1,I2,I5}/ {I5} = 2/2 = 100%
- R6 is Selected.

In this way, We have found three strong association rules.

ข้อดีของ Apriori

- ช่วยให้ทราบพฤติกรรมของเป้าหมายได้ โดยการใช้อัลกอริทึมจัดการเชื่อมความสัมพันธ์ของเหตุการณ์ต่างๆ ที่เราต้องการหาความสัมพันธ์ของเป้าหมาย คัดกรองข้อมูลออกมาตามความสัมพันธ์ วิเคราะห์ข้อมูลมาจนมีความน่าเชื่อถือและนำไปใช้ได้จริง

Question & Answer

ข้อเสียของ Apriori

- อัลกอริทึม Apriori ถือเป็นอัลกอริทึมที่นิยมใช้ในการหากฎความสัมพันธ์ของข้อมูล แต่ถ้าฐานข้อมูลมีการเพิ่มข้อมูลเข้ามาหรือเกิดมีการเปลี่ยนแปลงข้อมูล อัลกอริทึม Apriori จะต้องนำข้อมูลทั้งหมดมารวมกันก่อน แล้วจึงจะสามารถนำข้อมูลทั้งหมดไปค้นหากฎความสัมพันธ์ใหม่ทั้งหมด โดยไม่สามารถนำกฎความสัมพันธ์ที่หาได้จากกลุ่มข้อมูลเก่าก่อนหน้ามาใช้ให้เกิดประโยชน์ได้ ทำให้เสียเวลาในการทำงานเพื่อค้นหากฎความสัมพันธ์ใหม่ทั้งหมด

Question & Answer

Exercise

1.จงใช้เทคนิค Association Rule Discovery ในการค้นหากฎความสัมพันธ์ของข้อมูลโดยในตารางที่ 1 คือ ตัวอย่างชุดข้อมูลการซื้อสินค้า ซึ่งคอลัมน์ TID เปรียบเสมือนตะกร้าที่ใส่สินค้าที่ซื้อในครั้งหนึ่งๆ และคอลัมน์ Items คือรายการสินค้าที่ซื้อพร้อมกันใน TID ใดๆ และตัวอักษร A,B,C,D และ E แทนชื่อสินค้าแต่ละชนิด โดยที่กำหนดค่าสนับสนุนขั้นต่ำ (Minimum Support) เท่ากับ 50% และค่าความมั่นใจขั้นต่ำ (Minimum Confidence) เท่ากับ 70%

ตารางที่ 1 ตัวอย่างข้อมูลรายการซื้อสินค้า

TID	Items
1	A C D
2	B C E
3	A B C E
4	B E
5	A B C E

2. กำหนด Minimum Support เท่ากับ 50% และ Minimum Confidence เท่ากับ 70% พิจารณา Transaction ที่กำหนดให้ต่อไปนี้

ID	Items bought
20010712001	{Apple Mac, Bubble Jet, CD Writer}
20010712002	{DVD Rom, CD Writer}
20010712003	{Graphic Card, DVD Rom, Bubble Jet, Film }
20010712004	{CD Writer, E-Business SW, Bubble Jet, DVD Rom, Film }

2.1 หา Frequent itemsets ทั้งหมดโดยใช้ Apriori Algorithm

2.2 จาก Frequent itemsets ที่ได้ในข้อ 2.1 ให้หากฎความสัมพันธ์ที่อยู่ในรูป

"buys(item1) and buys(item2) → buys(item3)" ที่ผ่านเกณฑ์ของ Minimum Support และ Minimum Confidence ที่กำหนด

3. กำหนด Minimum Support = 33.34% และกำหนด Minimum Confidence = 60% พิจารณา Transaction ที่กำหนดให้ต่อไปนี้

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

a) หา Frequent itemsets ทั้งหมดโดยใช้ Apriori Algorithm

b) จาก Frequent itemsets ที่ได้ในข้อ a. ให้หากฎความสัมพันธ์ที่อยู่ในรูป "buys(item1) → buys(item2)" ที่ผ่านเกณฑ์ของ Minimum Support และ Minimum Confidence ที่กำหนด