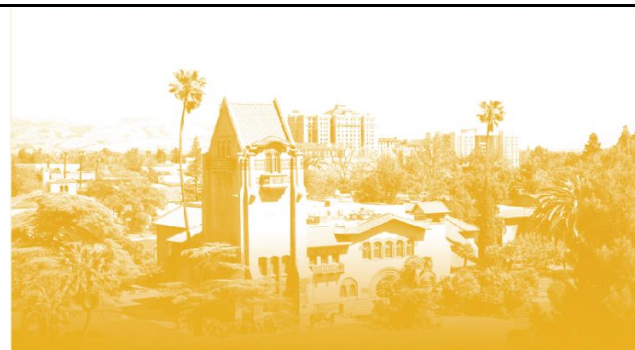


Set Data
Frequent Pattern Mining

1

Agenda

- Basic Concepts
- Frequent Pattern Mining
- Pattern Evaluation Methods

2

SAN JOSÉ STATE UNIVERSITY

Basics Concepts

3

SAN JOSÉ STATE UNIVERSITY

What is Set Data?

- Set Data: A collection of distinct objects, often represented in mathematical contexts. For instance, {apple, banana, cherry} is a set of fruits, and {1, 2, 3} is a set of numbers.
- Sets are useful because they allow us to group items and analyze their properties, like intersections and unions.
- Examples of Set Data:
  - {apple, banana, cherry}
  - {1, 2, 3}

4

SJSU SAN JOSÉ STATE UNIVERSITY

Set Data Datasets

- A data point corresponds to a set of items.
- Each data point is also called a transaction.

Transaction Dataset

Tid	Items bought
10	Beer, Nuts, Diaper
20	Beer, Coffee, Diaper
30	Beer, Diaper, Eggs
40	Nuts, Eggs, Milk
50	Nuts, Coffee, Diaper, Eggs, Milk

5

SJSU SAN JOSÉ STATE UNIVERSITY

What Is Frequent Pattern Mining?

- Frequent Pattern: a pattern (a set of items, subsequences, substructures, etc.) that occurs frequently in a data set.
  - First proposed by Agrawal, Imielinski, and Swami in1993
  - In the context of frequent itemsets and association rule mining
- Motivation: Finding inherent regularities in data
  - What products were often purchased together? → Beer and diapers?!
  - What are the subsequent purchases after buying a PC?
  - What kinds of DNA are sensitive to this new drug?
  - What's the next movie you will watch after watching a particular movie on Netflix?

6

SJSU SAN JOSÉ STATE UNIVERSITY

Importance of Pattern Mining

- Finding inherent regularities in a data set
- Foundation for many essential data mining tasks:
  - Association, correlation, and causality analysis
  - Mining sequential, structural (e.g., sub-graph) patterns
  - Pattern analysis in spatiotemporal, multimedia, time-series, and stream data
  - Classification: Discriminative pattern-based analysis
  - Cluster analysis: Pattern-based subspace clustering
- Broad applications

7

SJSU SAN JOSÉ STATE UNIVERSITY

Basic Concepts: Frequent Patterns

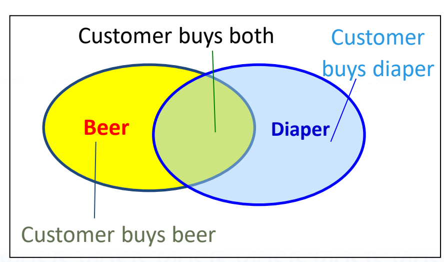
- Itemset: A set of one or more items $I = \{I_1, \ldots, I_N\}$
- k-itemset: $X = \{x_1, \ldots, x_k\}$

  e.g. {Beer, Nuts, Diaper} is a 3-itemset
- (*absolute*) *Support* of X, sup(X) : frequency or # of occurrences of an itemset X

  sup{Beer} = 3
  sup{Diaper} = 4
  sup{Beer, Diaper} = 3
  sup{Beer, Eggs} = 1
- (*relative*) *Support* of X, s(X) : fraction of transactions that contains X (i.e. the probability that a transaction contains X: P(X))

  s{Beer} = 3/5 = 60%
  s{Diaper} = 4/5 = 80%
  s{Beer, Eggs} = 1/5 = 20%

Tid	Items bought
10	Beer, Nuts, Diaper
20	Beer, Coffee, Diaper
30	Beer, Diaper, Eggs
40	Nuts, Eggs, Milk
50	Nuts, Coffee, Diaper, Eggs, Milk



8

SJSU SAN JOSÉ STATE UNIVERSITY

Basic Concepts: Frequent Itemsets (Patterns)

- An itemset/pattern X is frequent if sup(X) $\geq \sigma$ (*minsup* threshold)
- For the given 5-transaction dataset, take σ = 50%:
  - All the frequent 1-itemsets:
    sup(Beer): 3/5 (60%); sup(Nuts): 3/5 (60%);
    sup(Diaper): 4/5 (80%); sup(Eggs): 3/5 (60%)
  - All the frequent 2-itemsets:
    sup({Beer, Diaper}): 3/5 (60%)
  - All the frequent 3-itemsets: None
- Why do these itemsets form the complete set of frequent k-itemsets (patterns) for any k?
- What's the implication for a large dataset?

Tid	Items bought
10	Beer, Nuts, Diaper
20	Beer, Coffee, Diaper
30	Beer, Diaper, Eggs
40	Nuts, Eggs, Milk
50	Nuts, Coffee, Diaper, Eggs, Milk

9

SJSU SAN JOSÉ STATE UNIVERSITY

Basic Concepts: Association Rules

- An Association Rule is a rule of the form $X \rightarrow Y$ where:
  - X and Y are itemsets,
  - and $X \cap Y = \emptyset$
- Example:
  - {Diaper, Beer} → {Nuts}
  - {Diaper, Coffee} → {Nuts}
  - {Diaper} → {Beer}
- How strong is this rule? ➔ Need to quantify them using support and confidence
  - Measuring association rule between 2 itemsets: (Notation: X → Y [s, c])

$X \rightarrow Y$ [*support* = 20%, *confidence* = 60%]

Tid	Items bought
10	Beer, Nuts, Diaper
20	Beer, Coffee, Diaper
30	Beer, Diaper, Eggs
40	Nuts, Eggs, Milk
50	Nuts, Coffee, Diaper, Eggs, Milk

10

Support of an Association Rule

Tid	Items bought
10	Beer, Nuts, Diaper
20	Beer, Coffee, Diaper
30	Beer, Diaper, Eggs
40	Nuts, Eggs, Milk
50	Nuts, Coffee, Diaper, Eggs, Milk

- Support of a rule $X \rightarrow Y$:

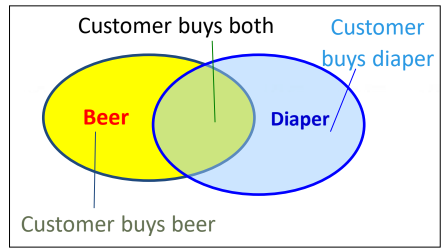
$$\text{sup}(X \rightarrow Y) = \text{sup}(X \cup Y) \;/\; \underset{\text{\# of transactions}}{D} = \text{P}(X \cup Y)$$

- The probability that a transaction contains $X \cup Y$
- Example:

  sup({Diaper} → {Beer}) = sup({Diaper, Beer})/D = 3/5 (60%)

  sup({Diaper, Coffee} → {Nuts}) = sup({Diaper, Beer, Nuts})/D
  = 2/5 (40%)

  sup({Diaper, Nuts} → {Milk}) = sup({Diaper, Nuts, Milk})/D
  = 1/5 (20%)



{Beer} ∪ {Diaper} = {Beer, Diaper}

11

Confidence of an Association Rule

Tid	Items bought
10	Beer, Nuts, Diaper
20	Beer, Coffee, Diaper
30	Beer, Diaper, Eggs
40	Nuts, Eggs, Milk
50	Nuts, Coffee, Diaper, Eggs, Milk

- Confidence of a rule $X \rightarrow Y$:

$$\text{conf}(X \rightarrow Y) = \text{sup}(X \cup Y) \;/\; \text{sup}(X) = \text{P}(Y \mid X)$$

- The conditional probability that a transaction containing X also contains Y
- Example:

  conf({Diaper} → {Beer}) = sup({Diaper, Beer} / sup({Diaper}) = 3/4 (75%)

  conf({Beer} → {Diaper}) = sup({Diaper, Beer} / sup({Beer}) = 3/3 (100%)

  conf({Beer, Diaper} → {Coffee}) = sup({Beer, Diaper, Coffee} / sup({Beer, Diaper}) = 1/3 (33.3%)

12

SJSU SAN JOSÉ STATE UNIVERSITY

Association Rule Mining

- Given two thresholds: *minsup* ϵ [0,1], *minconf* ϵ [0,1]
  find all rules $X \rightarrow Y$ [sup, conf]
  such that, sup ≥ *minsup* and conf ≥ *minconf*
- Example: Let *minsup* = 50%, *minconf* = 50%
  - 1-itemsets: {Beer}: 3, {Nuts}: 3, {Diaper}: 4, {Eggs}: 3
  - 2-itemsets: {Beer, Diaper}: 3
  - Beer → Diaper (60%, 100%)
  - Diaper → Beer (60%, 75%)

Are these all rules?

Tid	Items bought
10	Beer, Nuts, Diaper
20	Beer, Coffee, Diaper
30	Beer, Diaper, Eggs
40	Nuts, Eggs, Milk
50	Nuts, Coffee, Diaper, Eggs, Milk

- Mining association rules and mining frequent patterns are very close problems.
- Scalable methods are needed for mining large datasets

13

SJSU SAN JOSÉ STATE UNIVERSITY

Support and Confidence of Association Rules

Support:
- measure how frequently an itemset $\{X \cup Y\}$ appears in the dataset.
- find patterns that are less likely to be random.
- reduce the number of patterns.
- make the algorithms more efficient.

Confidence:
- measure the strength of associations.
- obtain an estimation of the conditional probability $P(Y|X)$.

Warning: A strong association does not mean that there is causality!

14

SJSU SAN JOSÉ STATE UNIVERSITY

Computational Complexity of Frequent Pattern Mining

- A long pattern contains a combinatorial number of sub-patterns
- How many frequent itemsets does the following dataset contain?
  - T1: $\{a_1, a_2, \ldots, a_{50}\}$
    T2: $\{a_1, a_2, \ldots, a_{100}\}$
  - Assuming (absolute) *minsup* = 1:

    1-itemsets: $\{a_1\}$: 2, $\{a_2\}$: 2, …, $\{a_{50}\}$: 2, $\{a_{51}\}$: 1, …, $\{a_{100}\}$: 1,

    2-itemsets: $\{a_1, a_2\}$: 2, …, $\{a_1, a_{50}\}$: 2, $\{a_1, a_{51}\}$: 1 …, …, $\{a_{99}, a_{100}\}$: 1,

    …, …, …, …

    99-itemsets: $\{a_1, a_2, \ldots, a_{99}\}$: 1, …, $\{a_2, a_3, \ldots, a_{100}\}$: 1

    100-itemset: $\{a_1, a_2, \ldots, a_{100}\}$: 1
- The total number of frequent itemsets: $\binom{100}{1} + \binom{100}{2} + \binom{100}{3} + \cdots + \binom{100}{100} = 2^{100} - 1$

15

SJSU SAN JOSÉ STATE UNIVERSITY

Computational Complexity of Frequent Pattern Mining

- How many itemsets are potentially generated in the worst case?
- The number of frequent itemsets generated is sensitive to the minsupthreshold
- When minsup is low, there exists potentially an exponential number of frequent itemsets!
- The worst case scenario:

$$\binom{M}{1} + \binom{M}{1} + \cdots + \binom{M}{N}$$

distinct items

max length of transactions

16

SJSU SAN JOSÉ STATE UNIVERSITY

Example: Frequent Pattern Mining

- Given the following transaction dataset, find frequent itemsets with min threshold = 0.2

Transaction	Red	White	Blue	Orange	Green	Yellow
1	1	1	0	0	1	0
2	0	1	0	1	0	0
3	0	1	1	0	0	0
4	1	1	0	1	0	0
5	1	0	1	0	0	0
6	0	1	1	0	0	0
7	1	0	1	0	0	0
8	1	1	1	0	1	0
9	1	1	1	0	0	0
10	0	0	0	0	0	1

Item set	Support
Red	0.6
White	0.7
Blue	0.6
Orange	0.2
Green	0.2
Red, White	0.4
Red, Blue	0.4
Red, Green	0.2
White, Blue	0.4
White, Orange	0.2
White, Green	0.2
Red, White, Blue	0.2
Red, White, Green	0.2

17

SJSU SAN JOSÉ STATE UNIVERSITY

Example: Frequent Pattern Mining

- Find rules associated with minconf = 70%.

Item set	Support
Red	0.6
White	0.7
Blue	0.6
Orange	0.2
Green	0.2
Red, White	0.4
Red, Blue	0.4
Red, Green	0.2
White, Blue	0.4
White, Orange	0.2
White, Green	0.2
Red, White, Blue	0.2
Red, White, Green	0.2

Item set	Rule	Support (A and B)	Support (A)	Confidence
Red, Green	Green → Red	0.2	0.2	1.000
White, Orange	Orange → White	0.2	0.2	1.000
White, Green	Green → White	0.2	0.2	1.000
Red, White, Green	Red, Green → White	0.2	0.2	1.000
Red, White, Green	White, Green → Red	0.2	0.2	1.000
Red, White, Green	Green → Red, White	0.2	0.2	1.000

18

SJSU SAN JOSÉ STATE UNIVERSITY

Frequent Pattern Mining

19

SJSU SAN JOSÉ STATE UNIVERSITY

Various Scalable Frequent Itemset Mining Methods

- Apriori: A Candidate Generation-and-Test Approach
- FPGrowth: A Frequent Pattern-Growth Approach
- ECLAT: Frequent Pattern Mining with Vertical Data Format

20

SJSU SAN JOSÉ STATE UNIVERSITY

The Apriori Algorithm

Apriori is one of the most classic and influential algorithms in frequent pattern mining.

- Developed by R. Agrawal and R. Srikant in 1994.
- Revolutionized how frequent pattern mining for large datasets was approached.
- Employs a bottom-up search method, where frequent subsets are extended one item at a time (candidate generation), and groups of candidates are tested against the data.
  - First find the complete set of frequent k-itemsets
  - Then derive frequent (k+1)-itemset candidates
  - Scan dataset again to find true frequent (k+1)-itemsets

21

SJSU SAN JOSÉ STATE UNIVERSITY

Apriori Properties

- Property 1: Given two itemsets X and Y. If $X \subset Y$, then
  $\sup(Y) \le \sup(X)$

  Example:
  - The support of {Diaper} = 4
  - The support of {Diaper, Eggs} = 2
  - The support of {Diaper, Eggs, Milk} = 1

Tid	Items bought
10	Beer, Nuts, Diaper
20	Beer, Coffee, Diaper
30	Beer, Diaper, Eggs
40	Nuts, Eggs, Milk, Cream
50	Nuts, Coffee, Diaper, Eggs, Milk

- Property 2 (Pruning): If an itemset X is infrequent, then its superset Y ($X \subset Y$) is also infrequent and shouldn't be generated or tested.

  Example:
  - Consider {Cream , Milk}
  - Since {Cream} is infrequent ➔ {Cream , Milk} is also infrequent

22

SJSU SAN JOSÉ STATE UNIVERSITY

The Apriori Algorithm

- Step 1: Candidate Generation:
  - Start with identifying all individual items in the dataset that meet the min support threshold.
  - Combine these items to form item sets of increasing size.
- Step 2: Pruning:
  - After creating larger item sets, those that don't meet the min support threshold are pruned out.
  - This pruning step is based on the Apriori property #2 (all non-empty subsets of a frequent item set must also be frequent).
- Step 3: Frequent Item Set Generation:
  - Repeat steps 1 & 2 until no more candidate item sets can be generated.

23

SJSU SAN JOSÉ STATE UNIVERSITY

Example: The Apriori Algorithm

C_k: Candidate itemsets of size k
L_k: Frequent itemsets of size k

$Sup_{min} = 2$

1st Iteration

Tid	Items
10	A, C, D
20	B, C, E
30	A, B, C, E
40	B, E

1st scan →

C_1

Itemset	sup
{A}	2
{B}	3
{C}	3
{D}	1
{E}	3

→ L_1

Itemset	sup
{A}	2
{B}	3
{C}	3
{E}	3

2nd Iteration

Why no {A, D} or {B, D}?

C_2

Itemset
{A, B}
{A, C}
{A, E}
{B, C}
{B, E}
{C, E}

2nd scan → C_2

Itemset	sup
{A, B}	1
{A, C}	2
{A, E}	1
{B, C}	2
{B, E}	3
{C, E}	2

→ L_2

Itemset	sup
{A, C}	2
{B, C}	2
{B, E}	3
{C, E}	2

3rd Iteration

Why no {A, C, E}?

C_3

Itemset
{B, C, E}

3rd scan →

Itemset	sup
{B, C, E}	2

→ L_3

Itemset	sup
{B, C, E}	2

24

SJSU SAN JOSÉ STATE UNIVERSITY

Performance of Apriori Algorithm

Performance of the Apriori algorithms depend on several factors:

- ***Minsup***: The lower it is, the larger the search space and the # of itemsets will be.
- **# of items**
- **# of transactions** (records) or **size of dataset**
- **Average transaction/record length**

25

SJSU SAN JOSÉ STATE UNIVERSITY

Strengths and Weaknesses of Apriori

Strengths

- Simplicity: The algorithm is easy to understand and implement.
- Efficiency: Effective for datasets with a relatively small number of transactions and items.

Weaknesses

- Scalability: Can be slow and inefficient for very large datasets due to the need to generate and count candidate itemsets as well as repeated scan of whole dataset.
- Memory Usage: Requires substantial memory to store numerous candidate itemsets, especially in later iterations when item sets become larger.

26

SJSU SAN JOSÉ STATE UNIVERSITY

Improvements of the Apriori Method

- Major computational challenges
  - Multiple scans of the entire datasets
  - Huge # of candidates
  - Tedious workload of support counting for candidates
- Improving Apriori: General Ideas
  - Reduce # of scans of dataset ➔ Using partition approach (only need to scan twice)
  - Reduce # of candidates ➔ Hash-based techniques
  - Facilitate support counting of candidates
- FP-Growth can effectively address the multiple scans and candidate generation issues.

27

SJSU SAN JOSÉ STATE UNIVERSITY

Pattern Evaluation Methods

28

SJSU SAN JOSÉ STATE UNIVERSITY

Misleading Strong Association Rules

- Not all strong association rules are interesting:

	Basketball	Not basketball	Sum (row)
Cereal	2000	1750	3750
Not cereal	1000	250	1250
Sum(col.)	3000	2000	5000

- Should we target people who plays basketball for cereal?

play basketball ⇒ eat cereal [40%, 66.7%]

play basketball ⇒ don't eat cereal [20%, 33.33%]

- Confidence measure of a rule could be misleading (66.7%) but the overall probability of people eating cereal is 75% (> 66.7%)!!

29

SJSU SAN JOSÉ STATE UNIVERSITY

Other Pattern Evaluation Measures

From Association to Correlation:

- Lift (next)
- χ^2
- All_confidence: all_conf(A, B) = min{$P(A|B)$, $P(B|A)$}
- Max_confidence: max_conf(A, B) = max{$P(A|B)$, $P(B|A)$}
- Kulczynski: Kulc(A, B) = ($P(A|B)$ + $P(B|A)$) / 2
- Cosine: cosine(A, B) = $\sqrt{P(A|B) \text{ x } P(B|A)}$

30

SJSU SAN JOSÉ STATE UNIVERSITY

Lift of an Association Rule

- Lift of a rule $X \rightarrow Y$:

$$lift(X \rightarrow Y) = \frac{conf(X \rightarrow Y)}{sup(Y)} = \frac{P(Y|X)}{P(Y)} = \frac{P(X \cup Y)}{P(X)P(Y)}$$

- It's the ratio of $conf(X \rightarrow Y)$ to $P(Y)$
- Measures the performance of the association rule against the baseline $P(Y)$

  Lift > 1: Positively correlated between A and B.

  Lift = 1: A and B are independent.

  Lift < 1: Negatively correlated between A and B.

31

SJSU SAN JOSÉ STATE UNIVERSITY

Example: Correlation Using Lift

- Using lift to evaluate the correlation between playing basketball and eating cereal etc:

	Basketball	Not basketball	Sum (row)
Cereal	2000	1750	3750
Not cereal	1000	250	1250
Sum(col.)	3000	2000	5000

$$lift(B \rightarrow C) = \frac{P(C|B)}{P(C)} = \frac{P(B \cup C)}{P(B)P(C)}$$

$$lift(B \rightarrow C) = \frac{P(B \cup C)}{P(B)P(C)} = \frac{\left(\frac{2000}{5000}\right)}{\left(\frac{3000}{5000}\right)\left(\frac{3750}{5000}\right)} = 0.89$$ ➡ negatively correlated!!

$$lift(B \rightarrow \bar{C}) = \frac{P(B \cup \bar{C})}{P(B)P(\bar{C})} = \frac{\left(\frac{1000}{5000}\right)}{\left(\frac{3000}{5000}\right)\left(\frac{1250}{5000}\right)} = 1.33$$ ➡ positively correlated!!

32

SJSU SAN JOSÉ STATE UNIVERSITY

Example Revisit

- Use lift to evaluate the rules:

Item set	Support
Red	0.6
White	0.7
Blue	0.6
Orange	0.2
Green	0.2
Red, White	0.4
Red, Blue	0.4
Red, Green	0.2
White, Blue	0.4
White, Orange	0.2
White, Green	0.2
Red, White, Blue	0.2
Red, White, Green	0.2

Item set	Rules	Support (A and B)	Support (A)	Confidence	Lift
Red, Green	Green → Red	0.2	0.2	1.000	1.667
White, Orange	Orange → White	0.2	0.2	1.000	1.429
White, Green	Green → White	0.2	0.2	1.000	1.429
Red, White, Green	Red, Green → White	0.2	0.2	1.000	1.429
Red, White, Green	White, Green → Red	0.2	0.2	1.000	1.667
Red, White, Green	Green → Red, White	0.2	0.2	1.000	2.500

33

SJSU SAN JOSÉ STATE UNIVERSITY

Summary

- Basic concepts:
  - frequent pattern, support, confidence and association rules
- Scalable frequent pattern mining methods
  - Apriori
- Which patterns are interesting?
  - Pattern evaluation methods such as lift etc

34