2.2 Frequent Pattern Mining - Association Rules

Association Rules in Action

Motivation

Originally stated in the context of **Market Basket Analysis**

- Data: set of items bought by costumers, (**transactions**)
- Find unexpected associations between sets of items using frequency of sets of items
- discovered sets of items: **frequent items** or **frequent patterns**

Actionable Knowledge

Shop Layout

Possible actions from rule {A1, A4} → {A6}

- Sell the A1, A4, A6 together (pack)
- Place article A6 next to articles A1, A4
- Offer a discount coupon for A6 in articles A1, A4
- Place a competitor of A6 next to A1, A4 (brand protection)

Cross-Selling

Steps:

- Client puts article A in basket
- Shop knows rule A → B
- Rule has enough confidence (> 20%)
- Shop tells client he may be interested in B
- Client decides whether to buy B or not

Notes:

- rules are discovered from business records
- discovery (mining) can be made offline
- use of rules can be made online

Text Mining

Each document is treated as a "bag" of terms and keywords

- **Goal**: identify co-occurring terms and keywords

Health

- Rules obtained from the patient's records
- We record the observations for each visit
- A set of observations may fire a rule
- Not necessarily causal

Web Usage Analysis

Usage patterns:

- Most visited pages
- Frequent page sets
- Pages associated to users
- Seasonal effects
- Cross-preferences

Association Rules Basic Concepts

- **Support**: measures the importance of a set
- **Confidence**: measures the strength of the rule

Mining Association Rules

Given:

- dataset of transactions D
- minimal support $minsup$
- minimal confidence $minconf$
  Obtain:
- all association rules $X \to Y (s = Sup, c = Conf)$ such that $Sup \geq minsup$ and $Conf \geq minconf$

Apriori Algorithm

1. **Frequent itemset generation**: itemsets with $support \geq minsup$
2. **Rule Generation**: generate all confident association rules from the frequent itemsets

- **Problem**: There is a very large number of candidate frequent itemsets
- **Downward Closure Property**
  - every subset of a frequent itemset must also be frequent

    - thus every superset of an infrequent itemset is also infrequent
- **Apriori Pruning Principle**: if an itemset is below the minimal support, discard all its supersets

Step 1 - Identifying Frequent Itemsets

- **Candidate generation** (Self-Join step): generates new candidates k-itemsets based on the frequent (k-1)-itemsets in the previous iteration
- **Candidate pruning** (Prune step): eliminates some candidate k-itemsets using the support-based pruning strategy

Step 2 - Rule Generation

- generate all non-empty subsets s of each frequent itemset I
- for each subset s compute the confidence of the rule $(I - s) \to s$
- select the rules whose confidence is higher than $minconf$
- **Note:** moving items from the antecedent to the consequent never changes support and never increases confidence.

Number of DB scans is n if the size of the largest frequent set is n or n-1.

Complexity Factors

- Nr. of items
- Nr. of transactions
- Minimal support
- Average size of transactions
- Nr. of frequent sets
- Average size of a frequent size
- Nr. of DB scans

Compact Representation of Itemsets

- The number of frequent itemsets produced from a transaction can be very large
- It is useful to identify a small representative set of itemsets from which all other frequent itemsets can be derived
- 2 such representations:
    - **Maximal**:
        - **Maximal frequent itemset**: frequent itemset for which none of its supersets is frequent
        - Can derive all frequent itemsets by computing all non-empty intersections
    - **Closed**:
        - **Closed frequent itemset**: frequent itemset that has no frequent supersets with the same support.

- Preserve the knowledge about the support values of all frequent itemsets

Reduce Rules

- Change parameters
- Restrict items
- Summarize techniques
- Filter rules

Improvement

Minimum difference between its confidence and the confidence of its immediate simplifications

Interesting Rule

- **Unexpected**: surprising to the user
- **Useful**: actionable
- **Subjective measures**: based on user's belief in the data
- **Objective measures**: based on facts, statistics and structures, independent of the domain considered
- Typically $A \rightarrow B$ is interesting if A and B are not statistically independent

LIFT

Ratio between confidence of the rule and the support of the itemset appearing in the consequent

- lift = 1; A and B are independent
- lift < 1; A and B are negatively correlated
- lift > 1; A and B are positively correlated

Conviction

Ratio between:

- the expected frequency that A occurs without B, if A and B were independent

- the observed frequency that the rule makes of incorrect predictions

- **High Conviction**: the consequent depends strongly on the antecedent

Improving Apriori

- Challenges of Frequent Pattern Mining
  - Multiple scans of transaction database

    - Huge number of candidates
    - Tedious workload of support counting for candidates
- **Ideas**:
    - Reduce number of transaction database scans
    - Shrink number of candidates (bottleneck of Apriori)
    - Facilitate support counting of candidates
- Methods to improve Apriori's efficiency
    - Partitioning
    - Sampling
    - Dynamic Itemset Counting
    - Frequent Pattern Projection and Growth (FP-Growth)