Association Analysis

- Mining for associations among items in a large database of transactions is an important data mining function.
- Association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases.
- Association rules are statements of the form {X1, X_2, …, Xn} => Y, meaning that if we find all of X_1, X_2, …… , Xn in the transaction then we have good chance of finding Y.
  Eg: The information that a customer who buys computer also tends to buy antivirus or pen drive.
- Association analysis mostly applied in the field of market basket analysis, web-based mining, intruder detection etc.

Market Basket Analysis

- Market basket analysis (also known as Affinity Analysis) is the study of items that are purchased or grouped together in a single transaction or multiple, sequential transactions.
- Understanding the relationships and the strength of those relationships is valuable information that can be used to make recommendations, cross-sell, up-sell, offer coupons, etc.
- A predictive market basket analysis can be used to identify sets of products/services purchased/events) that generally occur in sequence or something of interest to direct marketers.
- Advanced Market Basket Analysis provides an excellent way to get to know the customer and understand the different behaviors that can be leveraged to provide better assortment, design a better plan and devise more attractive promotions that can lead to more sales and profits.
- The analysis can be applied in various ways:
  - Develop combo offers based on products sold together.
  - Organize and place associated products/categories nearby inside a store.
  - Determine the layout of the catalog of an ecommerce site.
  - Control inventory based on product demands and what products sell together.

- Support of a product or group of products indicates the popularity of the product or group of products in the transaction set. Higher the support, more popular is the product or product bundle. This measure can help in identifying selling strategy of the store. Eg: if Barbie dolls have a higher support then they can be attractively priced to attract traffic to a store.
- Confidence can be used for product placement strategy and increasing profitability. Place high-margin items with associated high selling items. If Market Basket Analysis indicates

that customers who bought high selling Barbie dolls also bought high-margin candies, then candies should be placed near Barbie dolls.

- Lift indicates the strength of an association rule over the random co-occurrence of Item A and Item B, given their individual support. Lift provides information about the change in probability of Item A in presence of Item B. Lift values greater than 1.0 indicate that transactions containing Item B tend to contain Item A more often than transactions that do not contain Item B.
- In order to gain better insights, Market Basket Analysis can based on
    - Weekend vs weekday sales
    - Month beginning vs month-end sales
    - Different seasons of the year
    - Different stores
    - Different customer profiles

- Although Market Basket Analysis mostly applied for shopping carts and supermarket shoppers, there are many other areas in which it can be applied such as:

For a financial services company

- Analysis of credit and debit card purchases.
- Analysis of cheque payments made.
- Analysis of services/products taken e.g. a customer who has taken executive credit card is also likely to take personal loan.

For a telecom operator

- Analysis of telephone calling patterns.
- Analysis of value-add services taken together.

Few terminologies used in association analysis

Support: The support of an association pattern is the percentage of task-relevant data transaction for which the pattern is true.

Support (A): Number of tuples containing A / Total number of tuples
Support (A = > B): Number of tuples containing A and B / Total number of tuples

- If minsup is set too high, we could miss itemsets involving interesting rare items (e.g., expensive products)
- If minsup is set too low, it is computationally expensive and the number of itemsets is very large

Confidence: Confidence is defined as the measure of certainty or trustworthiness associated with each discovered pattern.

Confidence (A = > B): Number of tuples containing A and B / Total count of A

Itemset

– A collection of one or more items. Example: {Milk, Bread, Diaper}

– An itemset that contains k items is called k-itemset.

Frequent Itemset

– An itemset whose support is greater than or equal to a minimum support threshold.

Association Rule

– An implication expression of the form X => Y, where X and Y are itemsets.

Example: {Milk, Diaper} => {Beer}

Maximal Frequent Itemset:

- An itemset is maximal if none of its immediate supersets is frequent.

Closed Itemset:

- An itemset is closed if none of its immediate supersets has same support as of the itmeset.

Lift

- Lift is a measure of the performance of a targeting model (association rule) at predicting or classifying cases as having an enhanced response with respect to the population as a whole, measured against a random choice targeting model.
- Lift can be found by dividing the confidence by the unconditional probability of the consequent, or by dividing the support by the probability of the antecedent times the probability of the consequent.
- If some rule had a lift of 1, it would imply that the probability of occurrence of the antecedent and that of the consequent are independent of each other. When two events are independent of each other, no rule can be drawn involving those two events.
- If the lift is > 1, that lets us know the degree to which those two occurrences are dependent on one another, and makes those rules potentially useful for predicting the consequent in future data sets.
- Lift = P(Y | X) /P(Y)

Association Rules Mining

- Given a set of transactions T, the goal of association rule mining is to find all rules having support ≥ minsup threshold and confidence ≥ minconf threshold.
- Some of approaches for association rules mining are:

Brute- Force Approach

- List all possible association rules.
- Compute the support and confidence for each rule.
- Prune rules that fail to minimum support and minimum confidence level.

This approach is computationally very expensive.

Frequent Itemset Generation Strategies

- Reduce the number of candidates (M): For complete search, M=2d. Use pruning techniques to reduce M.
- Reduce the number of transactions (N): Reduce size of N as the size of itemset increases.
- Reduce the number of comparisons (NM): Use efficient data structures to store the candidates or transactions. No need to match every candidate against every transaction.

Apriori Approach

- Apriori approach is two step approach: Frequent item generation and Rules generation
- Based on apriori principal

Apriori Principle:

- Supersets of non-frequent item are also non-frequent. Or, If an itemset is frequent, then all of its subset also be frequent.
- Apriori algorithm is an influential algorithm for mining frequent itemset.
- It use a level-wise search, k-itemsets are used to explore k+1 itemsets.
- At first, the set of frequent itemset is found and used to generate to frequent itemset at next level and so on.

Apriori Algorithm:

- Read the transaction database and get support for each itemset, compare the support with minimum support to generate frequent itemset at level 1.
- Use join to generate a set of candidate k-itmesets at next level.
- Generate frequent ietmsets at next level using minimum support.
- Repeat step 2 and 3 until no frequent itme sets can be generated.
- Generate rules form frequent itemsets from level 2 onwards using minimum confidence.

This approach has faster than Brute-Force approach but still has higher computational complexity.

Example: Refer class note

Reducing Number of Comparisons

Candidate counting:

- Scan the database of transactions to determine the support of each candidate itemset.
- To reduce the number of comparisons, store the candidates in a hash structure
- Instead of matching each transaction against every candidate, match it against candidates contained in the hashed buckets

Hash Table

- A hash table (hash map) is a data structure used to implement an associative array, a structure that can map keys to values.
- A hash table uses a hash function to compute an index into an array of buckets or slots, from which the correct value can be found.
- Max leaf size: max number of itemsets stored in a leaf node, if number of candidate itemsets exceeds max leaf size, split the node.

Factors Affecting Complexity

i. ***Choice of minimum support threshold:*** Lowering support threshold results in more frequent itemsets.This may increase number of candidates and max length of frequent itemsets.

ii. ***Dimensionality (number of items) of the data set:*** More space is needed to store support count of each item. If number of frequent items increases, both computation and I/O costs may also increase.

iii. ***Size of database:*** Since Apriori makes multiple passes, run time of algorithm may increase with number of transactions.

iv. ***Average transaction width:*** Transaction width increases with denser data sets. This may increase max length of frequent itemsets and traversals of hash tree (number of subsets in a transaction increases with its width)

Frequent Pattern (FP) Growth Method

- Mining frequent itmesets without candidate generation.
- It is a divide and conquers strategy.
- It compress the database representing frequent items into a frequent –pattern tree (FP-Tree), which retains the itemset association information.
- Divides the compressed database into a set of conditional databases, each associated with one frequent item or pattern fragment and then mines each such database separately.
- FP-Growth method transforms the problem of finding long frequent patterns to searching for shorter ones recursively and then concatenating the suffix.
- It uses least frequent items as suffix .
- Adv: Reduce search cost, has good selectivity, faster than apriori.

- Disadv: When the database is large, it is sometimes unrealistic toconstruct a man memory based FP-tree.

FP-Tree algorithm

- Create root node of tree, labeled with null.
- Scan the transactional database.
- The items in each transaction are processed in sorted order (Descending) and branch is created for each transaction.

FP-Tree algorithm

- Start from each frequent length pattern as an initial suffix pattern.
- Construct conditional pattern base. (Pattern base is a sub database which consists of the set of prefix paths in the FP-tree co-occurring with suffix pattern.
- Construct its FP-tree and perform mining recursively on such a tree

Example: Refer class note

Categorical data

- Categorical data is a statistical data type consisting of categorical variables, used for observed data whose value is one of a fixed number of nominal categories.
- More specifically, categorical data may derive from either or both of observations made of qualitative data, where the observations are summarized as counts or cross tabulations, or of quantitative data.
- Observations might be directly observed counts of events happening or they might counts of values that occur within given intervals.
- Often, purely categorical data are summarized in the form of a contingency table.
- However, particularly when considering data analysis, it is common to use the term "categorical data" to apply to data sets that, while containing some categorical variables, may also contain non-categorical variables.

Potential Issues

- ***What if attribute has many possible values:*** Example: attribute country has more than 200 possible values. Many of the attribute values may have very low support.

  Potential solution: Aggregate the low-support attributes values.

- ***What if distribution of attribute values is highly skewed:*** Example: 95% of the visitors have Buy = No. Most of the items will be associated with (Buy=No) item

  Potential solution: drop the highly frequent items

Handling Categorical Attributes

- Transform categorical attribute into asymmetric binary variables. i.e If the outcomes of a binary variable are not equally important.
- Introduce a new "item" for each distinct attribute- value pair.

Sequential Pattern

- Mining of frequently occurring ordered events or subsequences as patterns. Eg: web sequence, book issued in library etc.
- Used mostly in marketing, customer analysis, prediction modeling.
- A sequence is an ordered list of events where an item can occur at most in an event of a sequence but can occur multiple times in different events of a sequence.
- Given a set of sequences, where each sequence consists of a list of events or elements and each event consists of set of items, given a minimum support threshold, sequential pattern mining finds all frequent subsequences.
- Sequence with minimum support is called frequent sequence or sequential pattern.
- A sequential pattern with length 'l' is called an l-pattern sequential pattern.
- Sequential pattern is computationally challenging because such mining may generate combinationally explosive number of intermediate subsequences.
- For efficient and scalable sequential pattern mining two common approaches are:

    i. Mining the full set of sequential patterns
    ii. Mining only the set of closed sequential pattern

- A sequence database is a set of tuples with sequence_ID and sequences. Eg:

Sequence_ID	Sequence
1	{(a, (a,b,c), (a,c), (b,c)}
2	{(a,b,c), (a,d),e,(d,e)}
3	{(c,d), (a,d,e),e}
4	{ (e,f,),d,(a,b,c),f]

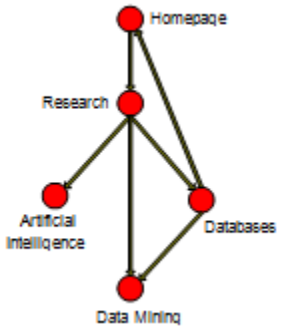
Sub-graph Patterns

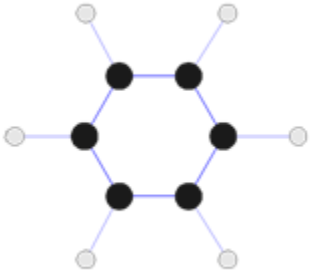
- It finds characteristics sub-graphs within the network.
- It is a form of graph search.
- Given a labeled graph data set, D = {G1, G2, ……..,Gn}, a frequent graph has minimum support not less than minimum threshold support.
- Frequent sub-graph pattern can be discovered by generating frequent substructures candidate and hence check the frequency of each candidate.
- Apriori methos and frequent –growth are tow common basic methods for finding frequent sub-graph
- Extend association rule mining to finding frequent subgraphs

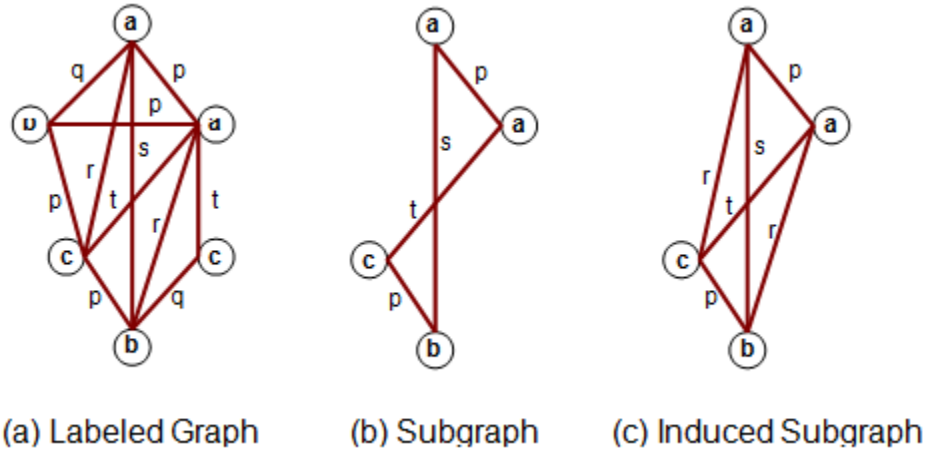
- Useful for Web Mining, computational chemistry, bioinformatics, spatial data sets, etc

  Eg::

  Chemical Structure, Geographical Nodes







Challenges

- Node may contain duplicate labels.
- How to define support and confidence?
- Additional constraints imposed by pattern structure
  - Support and confidence are not the only constraints
  - Assumption: frequent subgraphs must be connected

****Apriori-like* approach**:

– Use frequent k-subgraphs to generate frequent (k+1) subgraphs

‹

What is frequent-pattern mining in Data Streams?

- Frequent-pattern mining finds a set of patterns that occur frequently in a data set, where a pattern can be a set of items (called an itemset), a subsequence, or a substructure.
- A pattern is considered frequent if its count satisfies a minimum support. Scalable methods for mining frequent patterns have been extensively studied for static data sets.
- Challenges in mining data streams:

    - Many existing frequent-pattern mining algorithms require the system to scan the whole data set more than once, but this is unrealistic for infinite data streams.
    - A frequent itemset can become infrequent as well. The number of infrequent itemsets is exponential and so it is impossible to keep track of all of them.