Definitions

Lecture module 2: Basic Concepts

Frequent pattern: a pattern (itemset in transaction databases) that occurs as often or more often than a minimum support level.

Support level: symbolized by *s* or *σ*, the minimum number of occurrences of a pattern for it to be considered frequent.

Absolute support: the number of occurrences of an itemset in a transaction database.

Relative support: the frequency of occurrences of an itemset in a transaction database. It is Number of transactions containing itemset/total number of transactions.

*k*-itemset: an itemset that contains k items.

confidence: a measure of the association of one item with another. *p*(*A*|*B*)=*p*(*A*∪*B*)*p*(*B*) The conditional probability that a transaction containing B will also contain A.

Closed pattern: a pattern is closed if the itemset is frequent and there does not exist any superset with the same support. This is a lossless compression of the list of frequent itemsets.

Max pattern: a pattern is max if the itemset is frequent and there does not exist any superset that is frequent. This is a lossy compression of the list of frequent itemsets, because information about support is lost.

Lecture module 3: Mining Methods

Apriori algorithm: Insight is that once a pattern falls below the support threshold, none of its superpatterns can be above the support threshold.

ECLAT algorithm: Converts the data into a vertical format, where items are listed and associated transactions are treated as a set belonging to the item. Then forming the pattern can be done by grouping items that share transaction IDs.

FPgrowth algorithm: Converts a group of patterns into a tree structure.

Lecture module 4: Pattern Evaluation

Many measures exist for characterization of possible correlation between events A and B. In pattern evaluation context these are so called "interestingness measures", which are applied for example to itemsets of transactional data.

Note: almost there are two equivalent types of definition - in terms of set theory and in terms of the conditional probability. One can move from one to another using the following formula *p*(*A*|*B*)=*p*(*A*∪*B*)*p*(*B*) .

**Lift measure**:   
*lift*(*A*,*B*)≡*sup*(*A*∪*B*)*sup*(*A*)∗*sup*(*B*)=*p*(*A*∣∣*B*)*p*(*A*)

**Chi-squared measure**:   
*χ*2≡∑(*observed*−*expected*)2*expected*, where the summation of observed and expected values is over contingency table.

**AllConfidence measure**:   
*all*\_*conf*(*A*,*B*)≡*sup*(*A*∪*B*)*max*[*sup*(*A*),*sup*(*B*)]=*min*[*p*(*A*|*B*),*p*(*B*|*A*)]

**MaxConfidence measure**:   
*max*\_*conf*(*A*,*B*)≡*max*[*p*(*A*|*B*),*p*(*B*|*A*)]

**Cosine measure** (sometimes called "harmonized lift"):   
*cosine*(*A*,*B*)≡*sup*(*A*∪*B*)*sup*(*A*)∗*sup*(*B*√)=*p*(*A*|*B*)∗*p*(*B*|*A*)−−−−−−−−−−−−−√

**Kulczynski measure**:   
*Kulc*(*A*,*B*)≡12∗(*p*(*A*∪*B*)*p*(*A*)+*p*(*A*∪*B*)*p*(*B*))=*p*(*A*∣∣*B*)+*p*(*B*∣∣*A*)2

**Jaccard measure**:   
*Jaccard*(*A*,*B*)≡*p*(*A*∪*B*)*p*(*A*)+*p*(*B*)−*p*(*A*∪*B*)=1*p*(*A*∣∣*B*)+*p*(*B*∣∣*A*)*p*(*A*∣∣*B*)∗*p*(*B*∣∣*A*)−1

Lecture module 5: Mining Diverse Patterns

**MultiLevel patterns**

Level: Similar items form a higher level item, such as all types of milk. Lower levels might be specific types of milk (skim, 1%, 2%, whole)

**MultiDimensional patterns**

A dimension is a category along which data is collected. So there may be specific other things the store measures about the shopper in addition to the items in his/her basket.

**Compressed patterns**

*δ* is a measure of the closeness of two patterns. It is measured by the number of transactions they have in common divided by the total number of transactions belonging to either one of them.

**Colossal patterns**

In a transaction database that contains a frequent colossal pattern, most frequent small patterns will be subpatterns of the colossal pattern. Therefore you can fuse together small patterns to generate the large pattern.

Lecture module 6: Constraint-Based Mining

Pattern Space constraint: the pattern has to meet certain constraints, pruning takes place at the pattern level.

Data Space constraint: the transaction has to meet certain constraints, pruning takes place in the database by removing transactions.

Anti-monotonic constraint: A constraint that if violated for a set, will be violated for all supersets. Minimum support is an example of such a constraint.

Monotonic constraint: A constraint that if satisfied for a set, will be satisfied for all supersets.

Succinct constraint: A constraint that can be directly enforced by manipulating the data.

Convertible constraint: a constraint that can be converted to either anti-monotonic or monotonic by manipulating the data.

Lecture module 7: Sequential Pattern Mining

GSP (Generalized Sequential Patterns): an algorithm for handling sequential patterns that uses the insights of the Apriori algorithm for non-sequential patterns.

SPADE: an algorithm for handling sequential patterns that uses the insights of the ECLAT algorithm for non-sequential patterns.

PrefixSpan: an algorithm for handling sequential patterns that uses the insights of the FPGrowth algorithm for non-sequential patterns. Creates projected databases for restricted searching to grow the pattern (a projected database contains only transactions that have the element projected upon).

Prefix: a frequent element (single or pair of items). You project the database to only keep sequences where the prefix is the first item. Used with PrefixSpan.

Suffix: the sequences in the projected database after the prefix has been removed. Used with PrefixSpan.

Pseudo-Projection: Once the projected databases are small enough to fit into main memory, you can use pointers for converting the sequences to suffixes. This does not require disk access and speeds up the algorithm. This is used with PrefixSpan.

Lecture module 8: Graph Pattern Mining

There are two graph mining methods that are related to the Apriori pattern-mining algorithm. AGM grows by adding vertexes and FSG grows by adding edges. Adding edges is more efficient.

There is a method related to the pattern growth algorithms that is a graph-growing algorithm. gSpan.

SpiderMine is similar to PatternFusion for colossal patterns. The insight is the same, that you will find the subgraphs of frequent colossal graphs much more frequently than you will find independent small patterns.

Lecture module 9: Pattern-based Classification

CBA: Classification-Based Associations

CMAR: Classification based on Multiple Association Rules.

overfitting: This occurs when a set of rules performs well on the training data set but poorly on the test data set. This can be caused by a small training set, or noise in the data, or a systematic difference between the training data set and the test data set. This is a situation where pattern discovery compares directly to machine learning.

Lecture module 10: Frequent Pattern Mining for Text Data

corpus (corpora is plural): A curated collection of documents, sometimes pre-processed for text mining.

Latent Dirichlet Allocation (LDA): this is a probability distribution that reflects the assignment of words (unigrams, N-grams) to topics (main subjects of a piece of text). This probability distribution has nice mathematical properties that allow it to be updated with the addition of each N-gram.

N-gram: a number (*N*) of words.

collocation: a measure of how often words occur together in excess of their distribution in the document.

bag-of-words: A document is chopped up into its individual words and they are sampled randomly.

phrase intrusion: A method for testing the accuracy of text mining and topic classification. A phrase from another topic is inserted into the document and the algorithm is supposed to mark it as a different topic. The frequency of correct identification of the intruding phrase is a quality measure for the algorithm.

stop-word: a frequently occurring word that doesn't contain topic information (the, and, but, a).

token: a word after the grammatical markers have been removed (-ing, -ed, -s).

Discriminativeness (Purity): how purely a word is associated with a topic, and doesn't appear in other topics.

Lecture module 11: Advanced Topics

**Data Streams**

Error rate (*ϵ*): A small number (typically a fraction of the relative support minimum) that is controlled to make sure that frequent patterns are detected.

**Spatiotemporal and Trajectory Mining**

Flock: at least *m* objects within a circle of radius *r*, and move in the same direction.

Convoy: Use density-based clustering at each timestamp, not rigid circle.

Swarm: May not be together at each timestamp, but still moving in the same direction.

Periodic patterns:

**Software Bug Mining**

Example of finding copy-paste errors in software.

**Privacy Issues in Pattern Mining**

Three ways to preserve privacy

1) Input privacy (data hiding): obscure the actual inputs, for example by adding a random number to birthdates.

Data supplier anonymizes the data.

Third party anonymizes the data.

*k*-anonymity: Group the information into classes that have at least *k* members.

*l*-diversity: members of the group have *l* different classes.

2) Output privacy (knowledge hiding): Identify sensitive knowledge and don't disclose it.

3) Owner privacy: Hide the source of the data.

**Invisible Pattern Mining**

Pattern Mining as part of another function.