UNIT III INTRODUCTION TO DATA MINING 9

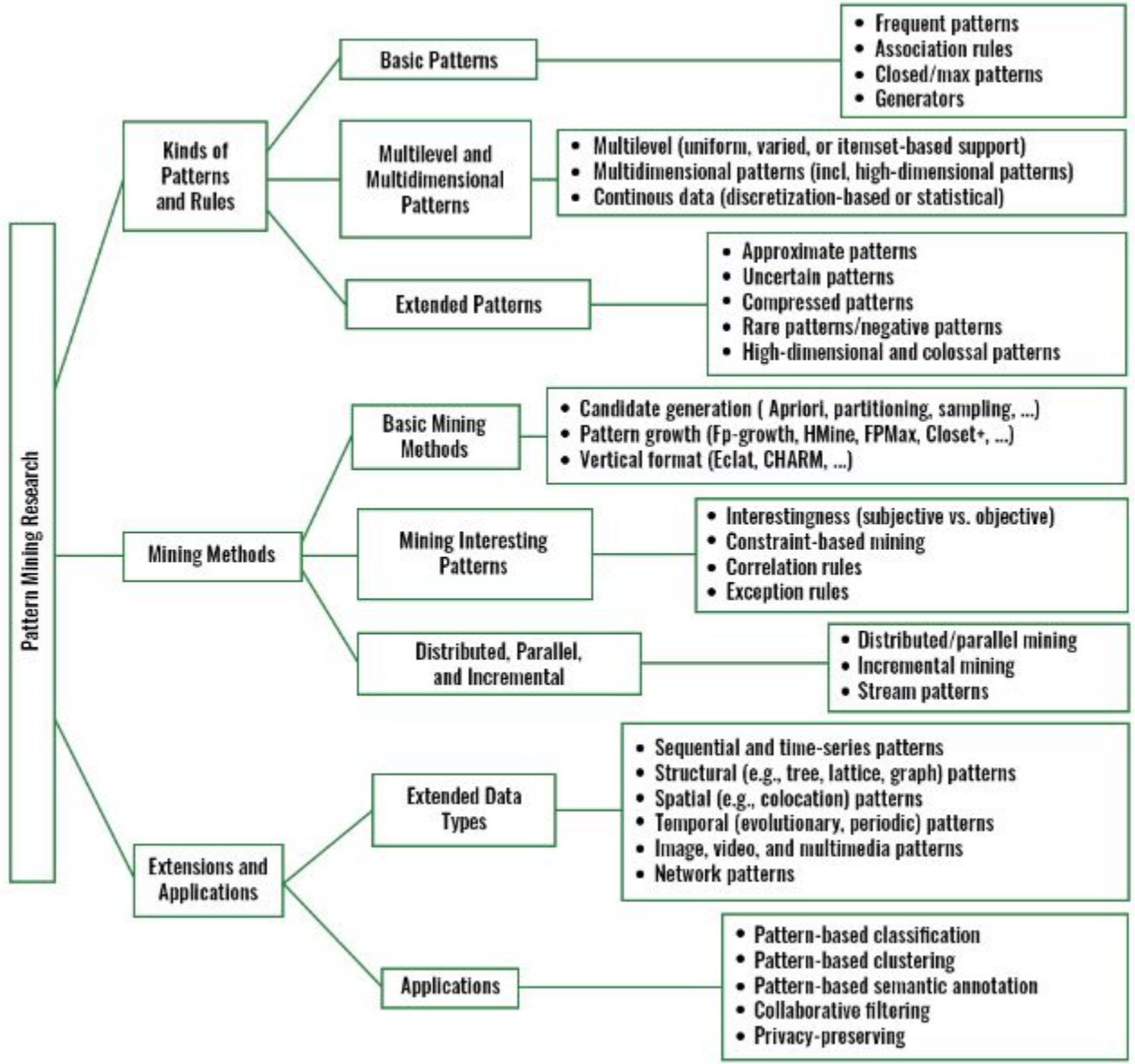
Data mining-KDD versus datamining, Stages of the Data Mining Process-task primitives, Data Mining Techniques -Data mining knowledge representation – Data mining query languages, Integration of a Data Mining System with a Data Warehouse – Issues, Data preprocessing – Data cleaning, Data transformation, Feature selection, Dimensionality reduction, Discretization and generating concept hierarchies-Mining frequent patterns- association-correlation

Mining frequent patterns

- The increasing power of computer technology creates a large amount of data and storage.
- Databases are increasing rapidly and in this computerized world everything is shifting online and data is increasing as a new currency.
- Data comes in different shapes and sizes and is collected in different ways.
- By using data mining there are many benefits it helps us to improve the particular process and in some cases, it costs saving or revenue generation.
- Data mining is commonly used to search a large amount of data for patterns and trends, and not only for searching it uses the data for further processes and develops actionable processes.

- Data mining has different types of patterns and **frequent pattern mining** is one of them.
- This concept was introduced for mining transaction databases.
- Frequent patterns are patterns(such as items, subsequences, or substructures) that appear frequently in the database.
- It is an analytical process that finds frequent patterns, associations, or causal structures from databases in various databases.
- This process aims to find the frequently occurring item in a transaction. By frequent patterns, we can identify strongly correlated items together and we can identify similar characteristics and associations among them.
- By doing frequent data mining we can go further for clustering

- Frequent data mining can be done by using association rules with particular algorithms eclat and apriori algorithms.
- Frequent pattern mining searches for recurring relationships in a data set. It also helps to find the inheritance regularities. to make fast processing software with a user interface and used for a long time without any error.


Pattern Mining Research
Kinds of Patterns and Rules
Basic Patterns
Frequent patterns
Association rules
Closed/max patterns
Generators
Multilevel and Multidimensional Patterns
Multilevel (uniform, varied, or itemset-based support)
Multidimensional patterns (incl. high-dimensional patterns)
Continous data (discretization-based or statistical)
Extended Patterns
Approximate patterns
Uncertain patterns
Compressed patterns
Rare patterns/negative patterns
High-dimensional and colossal patterns
Mining Methods
Basic Mining Methods
Candidate generation (Apriori, partitioning, sampling, ...)
Pattern growth (Fp-growth, HMine, FPMax, Closet+, ...)
Vertical format (Eclat, CHARM, ...)
Mining Interesting Patterns
Interestingness (subjective vs. objective)
Constraint-based mining
Correlation rules
Exception rules
Distributed, Parallel, and Incremental
Distributed/parallel mining
Incremental mining
Stream patterns
Extensions and Applications
Extended Data Types
Sequential and time-series patterns
Structural (e.g., tree, lattice, graph) patterns
Spatial (e.g., colocation) patterns
Temporal (evolutionary, periodic) patterns
Image, video, and multimedia patterns
Network patterns
Applications
Pattern-based classification
Pattern-based clustering
Pattern-based semantic annotation
Collaborative filtering
Privacy-preserving

Association Rule Mining:

- It is easy to find associations in frequent patterns:
- for each frequent pattern x for each subset y c x.
- calculate the support of y-> x – y.
- if it is greater than the threshold, keep the rule. There are two algorithms that support this lattice

1. Apriori algorithm
2. eclat algorithm

Apriori

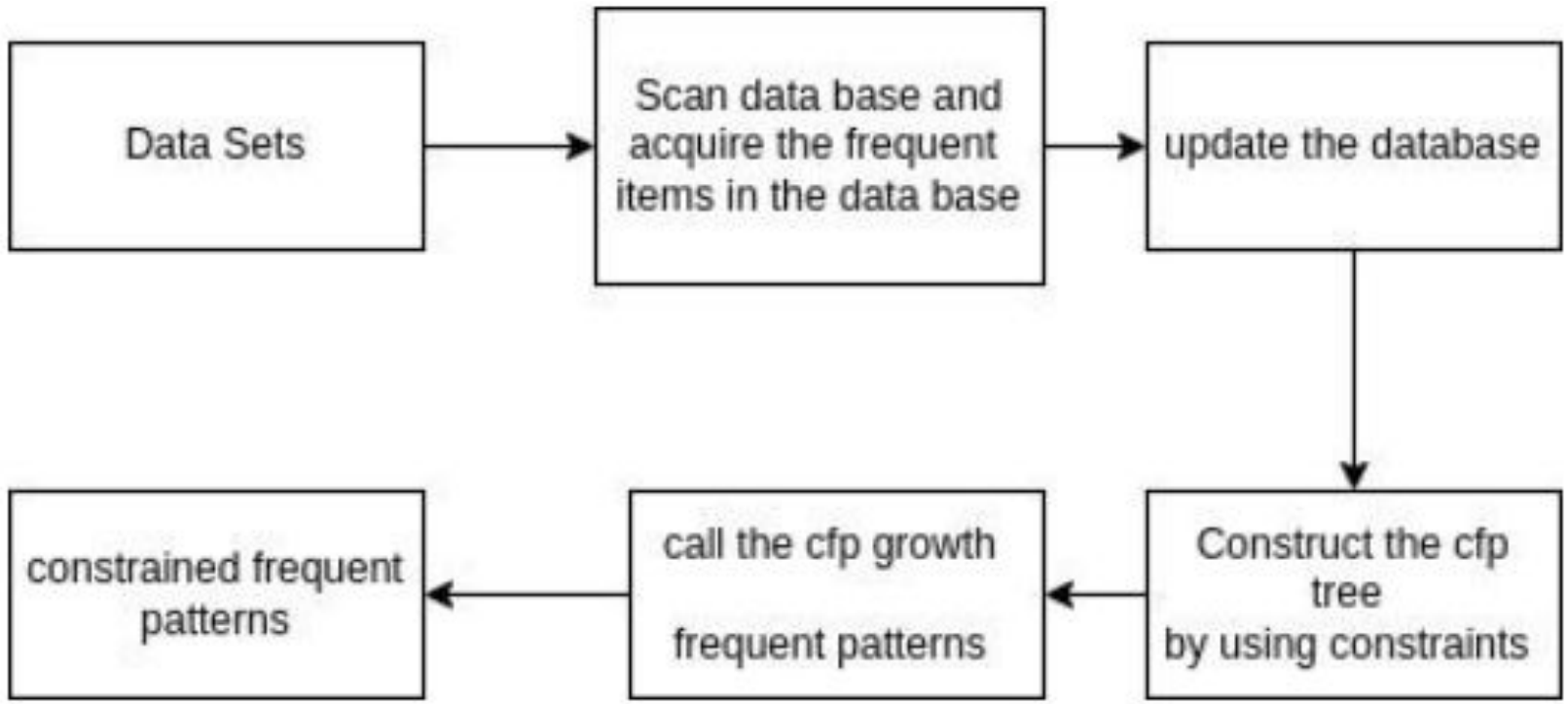
It performs “perfect” pruning of infrequent item sets.

It requires a lot of memory(all frequent item sets are represented) and support counting takes very long for large transactions. But this is not efficient in practice.

Eclat

It reduces memory requirements and is faster.

Its storage of transaction list.


Data Sets
Scan data base and acquire the frequent items in the data base
update the database
Construct the cfp tree by using constraints
call the cfp growth
frequent patterns
constrained frequent patterns

- The words support and confidence support the association rule.
- **Support:** how often a given rule in a database is mined? support the transaction contains x U y
- **Confidence:** the number of times the given rule in a practice is true. The conditional probability is a transaction having x as well as y.
- working principle (it is a simple point of scale application for any supermarket which has a good off-product scale)

- the product data will be entered into the database.
- the taxes and commissions are entered.
- the product will be purchased and it will be sent to the bill counter.
- the bill calculating operator will check the product with the bar code machine it will check and match the product in the database and then it will show the information of the product.
- the bill will be paid by the customer and he will receive the products.

- Tasks in the frequent pattern mining:
- Association
- **Cluster analysis:** frequent pattern-based clustering is well suited for high-dimensional data. by the extension of dimension the sub-space clustering occurs.
- **Data warehouse:** iceberg cube and cube gradient
- Broad applications
- There are some to improve the efficiency of the tasks.

- **Closed Pattern:**
- A frequent pattern, it meets the minimum support criteria. All super patterns of a closed pattern are less frequent than the closed pattern.
- **Max Pattern:**
- It also meets the minimum support criteria(like a closed pattern). All super patterns of a max pattern are not frequent patterns. both patterns generate fewer numbers of patterns so therefore they increase the efficiency of the task.
- **Applications of Frequent Pattern Mining:**
- basket data analysis, cross-marketing, catalog design, sale campaign analysis, web log analysis, and DNA sequence analysis.
- Issues of frequent pattern mining
- flexibility and reusability for creating frequent patterns
- most of the algorithms used for mining frequent item sets do not offer flexibility for resuing
- much research is needed to reduce the size of the derived patterns

- It is impossible to give complete coverage of this topic with the limited space and our limited knowledge. Frequent pattern mining has achieved tremendous progress and claimed a good set of applications. However in-depth research is required that the field may have a long-lasting and deep impact on data mining applications.