Pattern-Miner: Integrated Management and Mining over Data Mining Models

Evangelos E Kotsifakos, Irene Ntoutsi, Yannis Vrahoritis, Yannis Theodoridis
Patterns

Patterns:
- Compact, rich in semantics representation of raw data.
- (Ex.) Data Mining Patterns
  - Frequent itemsets & Association rules
  - Clusters
  - Decision trees
  - Time series
Pattern management issues

- Different and complex patterns
  - Data mining, image recognition, music, streams etc.
- Large databases
- Store, query, exchange, synchronize patterns
  - Handle, search discovered patterns – change patterns when data change

Pattern Miner offers:
- Future pattern exploitation
- Pattern Modeling
- Pattern Storage
- Pattern querying and retrieval
- Advanced pattern operations (comparison, monitoring)
Modeling Patterns (1)

- **Pattern-base**
  a collection of persistently stored patterns.

- **Pattern Base Mgmt System (PBMS)**
  - handles pattern-bases
  - efficiently supports pattern matching
  - executes pattern-related operations

*A PBMS treats patterns just like a DBMS treats raw data.*
Modeling Patterns (2)

Pattern type: \( pt = (n, ss, ds, ms, f) \)

- Name
- **Structure**
- **Source**
- Measures
- Formula

**Pattern**: An instance of Pattern Type

*Used for comparison and monitoring operations*
PatternMiner Architecture

- DM engine
- Pattern extraction
- Pattern storage
- Pattern querying
- Querying results
- Comparison results
- Patterns
- Pattern-Miner engine
- Meta-mining module
- Cluster Monitoring module
- (complex) Pattern extraction
- Meta-clustering results
- Pattern Base
- Pattern retrieval
Technologies

- WEKA data mining engine
- PMML XML standard for pattern modeling
- Berkeley XMLDB storage
- PANDA pattern comparison framework
PatternBase: Storage and Query

Pattern-Miner ver. 2.0

Pattern Base filename: Clustering

XQuery:
```
declare namespace a="http://www.dmg.org/PMML-3_2";
collection("Clustering.dbxml")
  [dbxml:metaData("dbxml:metaDataName"="C:\patternMiner\data_files\file001.art")]
  [a:PMML/a:ClusteringModel[@algorithmName="weka.clusterers.SimpleKMeans"]]
```

Execute XQuery

Finished
Pattern Comparison

- Comparison of patterns of the same pattern-type (i.e. itemsets, clusters etc)
- Different metrics to compare each pattern-type
- Additional user-defined metrics can be included
Pattern Similarity

- Pattern similarity is measured using the “structure” and “measure” pattern components, a matching and an aggregation functions.
Pattern Monitoring over time

- Discovers patterns changes over time
  - Based on the dataset (from which patterns have been extracted)
  - Based on the time (that patterns have been extracted)
  - Discovers new pattern appearances, disappearances, absorptions, splits
  - Currently works on Clusterings
Pattern Monitoring over time

Pattern-Miner ver. 2.0

Monitor...

Dataset: C:\patternMiner\data_files\file001.arff

Parameters for monitoring
Cluster's type: ClusteringCentroids
Survival threshold: 0.9
Split threshold: 0.2
Size threshold: 0.2
Location threshold: 0.2
Compactness threshold: 0.2

Parameters for clustering
Structure dissimilarity method: CohenDistance
Measure dissimilarity method: AbsoluteDistance
Distributions' aggregation function: avggood
Gatherer method: GathererZeroSimpleAvg

Executing...
Creating MON_km-F01-a-km-F01-b.xml...
Created MON_km-F01-a-km-F01-b.xml
Creating monitoring graph...
Finished
Pattern Monitoring over time
Conclusions

PatternMiner provides:

- Integrated pattern management environment
- Easy pattern exchange (XML model)
- Open source – expandable
  - Custom pattern types, measures and comparison techniques can be easily embedded.
Case scenarios

- cluster-based image retrieval
- pattern validation
- comparison of patterns extracted from different sites in a distributed environment setting
- Etc..
Future work

- Experimentation with various datasets
- Application in real scenarios
- Visualization module
- Incorporate more patternTypes
- Expand monitoring capabilities
References


