Mining Time-Series and Sequence Data

- Time-series database
  - Consists of sequences of values or events changing with time
  - Data is recorded at regular intervals
  - Characteristic time-series components
    - Trend, cycle, seasonal, irregular

- Applications
  - Financial: stock price, inflation
  - Biomedical: blood pressure
  - Meteorological: precipitation
Mining Time-Series and Sequence Data

A time series can be illustrated as a time-series graph which describes a point moving with the passage of time.

Categories of Time-Series Movements

- Long-term or trend movements (trend curve)
- Cyclic movements or cycle variations, e.g., business cycles
- Seasonal movements or seasonal variations
  - i.e., almost identical patterns that a time series appears to follow during corresponding months of successive years.
- Irregular or random movements
Estimation of Trend Curve

• The freehand method
  – Fit the curve by looking at the graph
  – Costly and barely reliable for large-scaled data mining

• The least-square method
  – Find the curve minimizing the sum of the squares of
    the deviation of points on the curve from the
    corresponding data points

• The moving-average method
  – Eliminate cyclic, seasonal and irregular patterns
  – Loss of end data
  – Sensitive to outliers

Discovery of Trend in Time-Series (1)

• Estimation of seasonal variations
  – Seasonal index
    • Set of numbers showing the relative values of a variable during the
      months of the year
    • E.g., if the sales during October, November, and December are
      80%, 120%, and 140% of the average monthly sales for the whole
      year, respectively, then 80, 120, and 140 are seasonal index
      numbers for these months
  – Deseasonalized data
    • Data adjusted for seasonal variations
    • E.g., divide the original monthly data by the seasonal index
      numbers for the corresponding months
**Discovery of Trend in Time-Series (2)**

- Estimation of cyclic variations
  - If (approximate) periodicity of cycles occurs, cyclic index can be constructed in much the same manner as seasonal indexes
- Estimation of irregular variations
  - By adjusting the data for trend, seasonal and cyclic variations
- With the systematic analysis of the trend, cyclic, seasonal, and irregular components, it is possible to make long- or short-term predictions with reasonable quality

**Similarity Search in Time-Series Analysis**

- Normal database query finds exact match
- Similarity search finds data sequences that differ only slightly from the given query sequence
- Two categories of similarity queries
  - Whole matching: find a sequence that is similar to the query sequence
  - Subsequence matching: find all pairs of similar sequences
- Typical Applications
  - Financial market
  - Market basket data analysis
  - Scientific databases
  - Medical diagnosis
Data transformation

- Many techniques for signal analysis require the data to be in the frequency domain
- Usually data-independent transformations are used
  - The transformation matrix is determined a priori
    - E.g., discrete Fourier transform (DFT), discrete wavelet transform (DWT)
  - The distance between two signals in the time domain is the same as their Euclidean distance in the frequency domain
  - DFT does a good job of concentrating energy in the first few coefficients
  - If we keep only first a few coefficients in DFT, we can compute the lower bounds of the actual distance

Multidimensional Indexing

- Multidimensional index
  - Constructed for efficient accessing using the first few Fourier coefficients
- Use the index can to retrieve the sequences that are at most a certain small distance away from the query sequence
- Perform post-processing by computing the actual distance between sequences in the time domain and discard any false matches
**Subsequence Matching**

- Break each sequence into a set of pieces of window with length \( w \)
- Extract the features of the subsequence inside the window
- Map each sequence to a “trail” in the feature space
- Divide the trail of each sequence into “subtrails” and represent each of them with minimum bounding rectangle
- Use a multipiece assembly algorithm to search for longer sequence matches

---

**Enhanced similarity search methods**

- Allow for gaps within a sequence or differences in offsets or amplitudes
- Normalize sequences with amplitude scaling and offset translation
- Two subsequences are considered similar if one lies within an envelope of \( \varepsilon \) width around the other, ignoring outliers
- Two sequences are said to be similar if they have enough non-overlapping time-ordered pairs of similar subsequences
- Parameters specified by a user or expert: sliding window size, width of an envelope for similarity, maximum gap, and matching fraction
Similar time series analysis

Steps for Performing a Similarity Search

• Atomic matching
  – Find all pairs of gap-free windows of a small length that are similar

• Window stitching
  – Stitch similar windows to form pairs of large similar subsequences allowing gaps between atomic matches

• Subsequence Ordering
  – Linearly order the subsequence matches to determine whether enough similar pieces exist
Query Languages for Time Sequences

- **Time-sequence query language**
  - Should be able to specify sophisticated queries like
    
    Find all of the sequences that are similar to some sequence in class A, but not similar to any sequence in class B
  - Should be able to support various kinds of queries: range queries, all-pair queries, and nearest neighbor queries

- **Shape definition language**
  - Allows users to define and query the overall shape of time sequences
  - Uses human readable series of sequence transitions or macros
  - Ignores the specific details
    - E.g., the pattern up, Up, UP can be used to describe increasing degrees of rising slopes
    - Macros: spike, valley, etc.
Sequential Pattern Mining

- Mining of frequently occurring patterns related to
time or other sequences
- Sequential pattern mining usually concentrate on symbolic patterns
- Examples
  - Renting “Star Wars”, then “Empire Strikes Back”,
    then “Return of the Jedi” in that order
  - Collection of ordered events within an interval
- Applications
  - Targeted marketing
  - Customer retention
  - Weather prediction

Mining Sequences (cont.)

Customer-sequence

<table>
<thead>
<tr>
<th>CustId</th>
<th>Video sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>{ (C), (H) }</td>
</tr>
<tr>
<td>2</td>
<td>{ (AB), (C), (DFG) }</td>
</tr>
<tr>
<td>3</td>
<td>{ (CEG) }</td>
</tr>
<tr>
<td>4</td>
<td>{ (C), (DG), (H) }</td>
</tr>
<tr>
<td>5</td>
<td>{ (H) }</td>
</tr>
</tbody>
</table>

Map Large Itemsets

<table>
<thead>
<tr>
<th>Large Itemsets</th>
<th>MappedID</th>
</tr>
</thead>
<tbody>
<tr>
<td>(C)</td>
<td>1</td>
</tr>
<tr>
<td>(D)</td>
<td>2</td>
</tr>
<tr>
<td>(G)</td>
<td>3</td>
</tr>
<tr>
<td>(DG)</td>
<td>4</td>
</tr>
<tr>
<td>(H)</td>
<td>5</td>
</tr>
</tbody>
</table>

Sequential patterns with support > 0.25

- { (C), (H) }
- { (C), (DG) }
Sequential pattern mining: Cases and Parameters

- Duration of a time sequence $T$
  - Sequential pattern mining can then be confined to the data within a specified duration
  - Ex. Subsequence corresponding to the year of 1999
  - Ex. Partitioned sequences, such as every year, or every week after stock crashes, or every two weeks before and after a volcano eruption
- Event folding window $w$
  - If $w = T$, time-insensitive frequent patterns are found
  - If $w = 0$ (no event sequence folding), sequential patterns are found where each event occurs at a distinct time instant
  - If $0 < w < T$, sequences occurring within the same period $w$ are folded in the analysis

Cases and Parameters (2)

- Time interval, $int$, between events in the discovered pattern
  - $int = 0$: no interval gap is allowed, i.e., only strictly consecutive sequences are found
    - Ex. “Find frequent patterns occurring in consecutive weeks”
  - $\min_int \leq int \leq \max_int$: find patterns that are separated by at least $\min_int$ but at most $\max_int$
    - Ex. “If a person rents movie A, it is likely she will rent movie B within 30 days” ($int \leq 30$)
  - $int = c \neq 0$: find patterns carrying an exact interval
    - Ex. “Every time when Dow Jones drops more than 5%, what will happen exactly two days later?” ($int = 2$)
Episodes and Sequential Pattern Mining Methods

- Other methods for specifying the kinds of patterns
  - Serial episodes: A → B
  - Parallel episodes: A & B
  - Regular expressions: (A | B)C*(D → E)

- Methods for sequential pattern mining
  - Variations of Apriori-like algorithms, e.g., GSP
  - Database projection-based pattern growth
    - Similar to the frequent pattern growth without candidate generation

Periodicity Analysis

- Periodicity is everywhere: tides, seasons, daily power consumption, etc.
- Full periodicity
  - Every point in time contributes (precisely or approximately) to the periodicity
- Partial periodicity: A more general notion
  - Only some segments contribute to the periodicity
    - Jim reads NY Times 7:00-7:30 am every week day
- Cyclic association rules
  - Associations which form cycles
- Methods
  - Full periodicity: FFT, other statistical analysis methods
  - Partial and cyclic periodicity: Variations of Apriori-like mining methods