Beyond the Crystal Ball –
Time Series Analysis and Forecasting
Claudio Hartmann – Lehrstuhl Datenbanken
System Architecture

Evolving Data-Driven Applications

Data Volume
Data Variety

Requirements

Throughput
Latency

Emerging Compute Unit Diversity

Emerging Network Diversity

Fully Connected

3D Mesh

Design and implement concepts to efficiently use these evolving hardware environment for database systems

Scalability

Reliability

Adaptability

Security

Emerging Memory Diversity

DRAM
HBM

Multi-Socket

FPGA
Vector
ASIC
NVRAM
Documents and Information Retrieval

- **Data Curation**
  - Column Specification
  - Semantic Normalization
  - Data Lake
  - WWW
  - semi-structured data

- **Hybrid DBMS/IR System**
  - Top-k Consistent Entity Augmentation System
  - Entity Augmentation Queries
  - Open-world SQL Queries
  - DrillBeyond

- **Multi-variant Query Processing**
  - Multi-variant Query Processing
  - Vector models
  - Retrofitting

- **Vector models**
  - *2vec

- **FREDDY**
  - retrofitting

- **REA**
  - REA
  - Entity Augmentation Queries

- **WWW**
  - WWW
  - open-world SQL Queries

- **Data Lake**
  - Data Lake
  - unstructured data

- **DeExcelarator**
  - DeExcelarator
  - from nation, customer, orders
  - where 
    - n.nationkey = o.nationkey 
    - o.o_id = c.custkey 
    - o._id > 10.0 
  - group by n.n nationkey, o._id, c.custkey 
  - order by o._id, c.custkey

- **FREDDY**
  - FREDDY
  - retrofitting

- **ANN Query Processing**
  - SELECT product_desc 
    - FROM products
  - ORDER BY cosine_similarity 
    - ('allergen', product_desc)
Data Analytics

Support/Extend Database Technology
- Sampling Techniques
- Integration of Machine Learning into RDBMS
- Cardinality estimation using Artificial Neural Networks
- Monitoring database health status

Working with Data
- Time series Forecasting
- Time series properties and generation
- Clustering with human feedback
Agenda

What is a time series?

**Health of a Database System**
- Monitoring and Target
- Techniques

**Feature-based Time Series Engineering**
- Time Series Features
- What-if analysis

**Forecasting Large-scale Time Series Data**
- Forecasting Process
- Big Data Implications
- CSAR
Terminology

**Time Series**

- Sequence of measure values
- Ordered by time
- Equidistant
  - Constant time distance between measure values
- Complete
  - No missing values
Health of a Database System
Health of a Database System

Monitoring
- Several measurements that serve as indicator for the health status of the database

Target
- Issue automatic warning for situations that would lead to customer complaints

Possible Health indicators
- CPU (user_busy, io_busy, system_busy, idle)
- user connections
- physical reads/writes, ...
Health of a Database System

**Threshold/Limits**
- Define upper threshold
- Issue warning when threshold is exceeded

**Problem**
- Manual work
- Proper definition of threshold
  - Too high → Warning too late
  - Too low → Waring when there is no issue
Health of a Database System

**Automatic threshold using IQR**
- Define threshold based on box plot statistics
- Issue warning when threshold is exceeded

**Problem**
- May detect normal states as outliers
Health of a Database System

Co-development of measurements
- Examine proportions of several measurements
- Issue warnings for abnormal proportions

Problem
- Know all effects in advance
- Lots of manual work
Health of a Database System

Structural Equation Modeling
- Model influences between system components
- Analysis of Latent Factors

Model Properties
- Item
  - Measurable variables
- Latent Factor
  - Not measurable variables
- Measurement model
  - Model connections between items and factors using covariance
Health of a Database System

**Structural Equation Modeling**
- Model influences between system components
- Analysis of Latent Factors

**Model Properties**
- Item
  - Measurable variables
- Latent Factor
  - Not measurable variables
- Measurement model
  - Model connections between items and factors using covariance
Feature-based Time Series Engineering
Process of Time Series Engineering

**Representation Engineering**
- Time Series Dataset
- Represent
- Representation Dataset
- Distance

**Data-mining Tasks**
- Generation
- Indexing
- Classification
- Clustering

Figures: Kegel et al., 2018; Shieh and Keogh, 2013; Fulcher et al., 2013; Wang et al., 2006
Example Time Series Features

**Distribution**
- Gaussian
- Mean = -0.5
- Skew = 0.3

**Correlation**
- ACF plot
- Lag: 0, 5, 10, 15, 20

**Components**

**Stationarity**

**Frequency Domain**

**Tests**
- $p \sim 0.94$
- $p < 2.2e-16$
- p-value of Normality test
Time Series Components

Decomposition

- Base: stationary part of the time series
- Trend: long-term change in the mean level
- Season: cyclical repeated behavior
- Residuals: unstructured information assumed to be random

Additive Composition

\[ x_t = \text{base}_t + \text{trend}_t + \text{season}_t + \text{res}_t \]

Often, base is part of the trend component!

Series

Base

Trend

Season

Residuals

Deterministic

Stochastic
Time Series Decomposition

Basic Idea

▪ Moving-average filter of continuous windows

\[ tr_t = \frac{x_t + x_{t-1} + x_{t-2} + \ldots + x_{t-N+1}}{N} \]

▪ Extraction of trend by windows that take into account the season length

▪ Extraction of season by averaging each time instance of the same seasonal position (all Mondays, all Tuesdays,...)

▪ Disadvantage: does not decompose the endpoints

Average Centering

▪ A technique needed if season length is even

▪ Take two moving averages and average their result

A season of length 7 such as:
- Monday (1)
- Tuesday (2)
- ...
Time Series Decomposition

**Retrieval of Trend**
- Moving-average filtering
- In case of even season length, centralize first

```
# Trend Retrieval
series = [1, 2, 3, 4, 5, 6, 7, 1, 2, 3, ...

# Centralize first

# Moving-average filtering
rep_len = 1,7

# Detrend series
detrend = series - trend

# Season Retrieval
season = figure - mean(figure)

# Residuals
residuals = series - season - trend
```
Feature-based Generation Method (FBG)

Idea
- Feature-based representation
- Recombination of deterministic features
- Simulation of residuals with statistical model

Use cases
- Anonymization of data sets
- Generate a data set that is closely related to the original data set but does not contain the actual data

Workflow

Given Dataset

Deterministic Features

Recombination

Stochastic Features

Statistical Model

Distribution

Autocorrelation of lag 1

Distance

Generated Dataset

Kegel et al., 2018
What-if Analysis

General Idea

- Represent time series by their features
- Generate a what-if scenario by setting factors that modify features

Series
Decomposition
Feature Extraction
What-If Scenario
Modification

Slope = 85

Slope = 100
Feature Extraction

Feature Space and Selected Time Series

Season Determination $R_{seas}^2$

$R_{seas}^2 = 1 - \frac{\text{var}(res_t)}{\text{var}(res_t + seas_t)}$

- **Weak Season**
- **Strong Season**
Feature Extraction

**Trend Determination** $R^2_{tr}$

- $R^2_{tr} = 1 - \frac{\text{var}(\text{rest})}{\text{var}(\text{rest}+\text{tr})}$

**Trend Slope** $\theta_2$

- Suppose a linear trend within STL trend:
  \[ tr_t = \theta_1 + \theta_2 \cdot l_t + \delta_t \]

**Trend Linearity** $R^2_{lin}$

- $R^2_{lin} = 1 - \frac{\text{var}(\delta_t)}{\text{var}(tr_t)}$
Component Modification

**Trend Determination Factor**
- Strengthen/weaken trend
  \[ tr_{t,f} = \theta_1 + f \cdot (\theta_2 \cdot l_t + \delta_t) \]
  \[ R_{tr}^2 = 1 - \frac{\text{var}(res_t)}{\text{var}(res_t + tr_{t,f})} \]

**Season Determination Factor**
- Strengthen/weaken season
  \[ seast_{k,k} = k \cdot seast_t \]
  \[ R_{seas}^2 = 1 - \frac{\text{var}(res_t)}{\text{var}(res_t + seast_{t,k})} \]
Forecasting Large-scale Time Series Data
Forecasting Process

1. Model Identification

<table>
<thead>
<tr>
<th>Method</th>
<th>Error</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARIMA</td>
<td>23.47</td>
<td>25.72</td>
</tr>
<tr>
<td>HoltWinters</td>
<td>18.38</td>
<td>20.43</td>
</tr>
<tr>
<td>VAR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MARS</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Optimize
- Predict
- Evaluate

2. Model Estimation

3. Forecast Calculation

4. Model Evaluation

5. Model Adaptation

\[
\text{SMAPE} = \frac{1}{n} \sum_{t=1}^{n} \frac{|x_t - \hat{x}_t|}{(x_t + \hat{x}_t)/2}
\]
Traditional Time Series Forecasting

Univariate Forecast Models

- ARIMA/Exponential Smoothing
- Focus on only one time series at a time
- Widely applied in many domains
- auto.ARIMA/ETS to properly configure the model for a given time series

\[
\hat{y}_{t+1} = c + \epsilon_t + \sum_{i=1}^{p} \phi_i y_{t-i+1} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j+1}
\]
Large-scale Time Series Data

Many and Long Series
- High number of monitored objects (Smart Meter in every household, sales of individual products)
- Fine monitoring granularity leads to very long time series histories

High Levels of Noise
- Time series on fine granularity tend to be very noisy

Missing values
- Missing values lead to inapplicability of all most conventional models
Cross-sectional Forecasting

Core Approach

- Represent a whole data set with one model
- Focus on cross-sections instead of the entire time series
- Model transition from one cross-section to the next one
- All time series with values in the blue cross-sections contribute to the model training
- The model represents the average transition of the entire data set

\[
\hat{y}_{t+1} = c + \phi_1 \cdot \hat{y}_t
\]
Cross-sectional Forecasting

Core Approach
- Represent a whole data set with one model
- Focus on cross-sections instead of the entire time series
- Model transition from one cross-section to the next one
- All time series with values in the blue cross-sections contribute to the model training
- The model represents the average transition of the entire data set

Model Application
- Assume the transition remains constant over seasons
- Apply model on the most current data
- Train a specific model for every transition in a season

\[
\hat{y}_{t+1} = c + \phi_1 \cdot \hat{y}_t
\]

Still misses adaptability!
CSAR – Autoregression

Non-seasonal Autoregression
- Model the dependency of future values of their direct predecessors

\[ \hat{y}_{t+1} = c + \phi_1 \cdot \hat{y}_t + \cdots + \phi_p \cdot \hat{y}_{t-p+1} \]

Seasonal Autoregression
- Model the dependency of future values of their seasonal predecessors

\[ \hat{y}_{t+1} = c \Phi_1 \cdot \hat{y}_{t-s+1} + \cdots + \Phi_p \cdot \hat{y}_{t-p+s+1} \]

Correction Terms
- Necessary if non-seasonal and seasonal components are combined
CSAR – Autoregression

Model Application
- On most recent data
- On all time series which have all necessary values

Behavioral Deviation
- Some time series do not follow the common behavior of the majority

\[
\hat{y}_{t+1} = c + \phi_1 \cdot \hat{y}_t + \phi_2 \cdot \hat{y}_{t-1} + \Phi_1 \cdot \hat{y}_{t-s+1} + (-\Phi_1 \phi_1) \cdot \hat{y}_{t-s} + (-\Phi_1 \phi_2) \cdot \hat{y}_{t-s-1}
\]
CSAR – Error Terms

Non-seasonal Error Terms
- Adjust forecasts according to systematic non-seasonal misprediction
- Deviation of individual time series from the general model for the data set

Seasonal Error Terms
- Adjust forecasts according to systematic seasonal misprediction
- Systematic deviation in the seasonal pattern

Missing Data
- If real value or forecast is missing, the value is ignored
CSAR – Model Components

Autoregression
- $p$ specifies number of non-seasonal AR terms
- $P$ specifies number of seasonal AR terms

Error Terms
- $f$ specifies number of non-seasonal error terms
- $F$ specifies number of seasonal error terms

Model Configuration
- Adaptation to data set specific characteristics
- Influence on forecast accuracy
- Influence on execution time
Step 1 – Base Models
- Compare basic model components
- Choose model with lowest error to identify most important model component

Step 2 – Search Around the best Model
- Based on result from Step 1
- Vary optimal model components by +/- 1
- Vary seasonal and non-seasonal components by +/- 1
- Invert the constant
- Repeat until an iteration returns no new best model
Forecast Accuracy

Experimental Set-Up
- Calculate long-range forecasts
  - Energy – 1 week (h=28)
  - Payment – 2 weeks (h=14)
- Calculate SAPE error between forecast and real time series values (Symmetric Absolute Percentage Error)

\[ SAPE = \frac{|y - \hat{y}|}{(|y| + |\hat{y}|)/2} \cdot 100 \]
Data Set Partitioning

Univariate vs. CSAR
- Opposite extremes
- Univariate one model per time series
- CSAR one model for all time series

Partitioning
- Split data set into several Partitions
- Create one CSAR model for each partition

Expectation
- Better representation of time series
- Higher forecast accuracy
Model Adviser

Model Suitability

- Different models work well with different time
- Some are designed for very specific use cases

Which one to use?

- Not an easy decision for non-experts
Topics for Student Theses

Student Theses at the Chair of Databases

https://wwwdb.inf.tu-dresden.de/study/theses/themen-fuer-arbeiten/

Themenschwerpunkte

Aktuell werden studentische Arbeiten in folgenden Themenschwerpunkten vergeben:

<table>
<thead>
<tr>
<th>Bereiche und aktuelle Angebote</th>
<th>Ansprechpartner</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DB-Systemarchitektur, skalierbare und sichere Datenverarbeitung, DB-Systeme auf moderner Hardware</strong></td>
<td>Dirk Habich</td>
</tr>
<tr>
<td>Eine Auswahl aktueller Themen.</td>
<td></td>
</tr>
<tr>
<td>- Benchmark Design für adaptive Datenbanksysteme</td>
<td></td>
</tr>
<tr>
<td>- Integration des ERIS Storage Systems in Apache Spark</td>
<td></td>
</tr>
<tr>
<td>- Implementierung und Optimierung eines Codesgenerators für Kompressionsalgorithmen</td>
<td></td>
</tr>
<tr>
<td>- Evaluierung der Intel SGX Erweiterung für eine sichere Datenverarbeitung</td>
<td></td>
</tr>
<tr>
<td>- Implementierung und Evaluation eines Storage Moduls für Graphdaten in ERIS</td>
<td></td>
</tr>
<tr>
<td><strong>Informationsextraktion, Information Retrieval, Machine Learning</strong></td>
<td>Maik Thiele</td>
</tr>
<tr>
<td>Eine Auswahl aktueller Themen.</td>
<td></td>
</tr>
<tr>
<td>- Word2Vec-Modell über Webtabellen</td>
<td></td>
</tr>
<tr>
<td>- Column-specification mit System T</td>
<td></td>
</tr>
<tr>
<td>- Zeitreihenanalyse und -prognose</td>
<td></td>
</tr>
</tbody>
</table>

Für weitere individuelle Angebote kontaktieren Sie bitte den zuständigen Ansprechpartner.