Ch. 12: Workload Forecasting

Kenneth Mitchell

School of Computing & Engineering,
University of Missouri-Kansas City,
Kansas City, MO 64110
Introduction

- Forecasting is the art and science of predicting future events
- Demand forecasting is essential for guaranteeing quality of service
- Critical for Web services
- Development of new and popular applications and services can create unanticipated service demands that can overwhelm a Web service
Demand Forecasting

- What will be the number of visits to the Web site?
- What will be the workload for the credit authorization service during the holiday season?
- How will the number of e-mail messages vary over the coming year?
- Web demand is characterized by a large portion of which is **unpredictable demand** (i.e., random fluctuations)
Forecasting strategy

- Can be grouped into categories
  1. Quantitative: Relies on the existence of historical data to estimate future values of workload parameters
  2. Qualitative: A subjective process, based on judgements, intuition, expert opinions, historical analogy, commercial knowledge, etc.

- Delphi method iteratively seeks consensus amongst experts

- Effects of P2P file sharing systems has no precedent and relies heavily on qualitative forecasting
Forecasting Strategy

Main activities involved in setting up a forecasting strategy:
- Forecasting objectives: How forecast will be used
- Qualitative forecasting: Directions of information technology
- Historical Data: business statistics or system logs
- Forecasting process: Forecasting span.
- Forecasting validation: Does model accurately predict
Forecasting Process

- Workload demands are too vague for end-user community
- Data come from existing system logs and from users
  - What will be the peak number of simultaneous users next year?
  - What will be the arrival rate of trivial transactions of the checking account application?
  - What will be the daily number of page views to the company’s Web site six months from now?
  - What will be the average size of e-mails by the end of next year?
- Answers to above questions should then be converted to typical workload parameters
Forecasting Models

Forecasting can be accomplished at many levels
- Business level
- User level
- Application level
- Protocol level
- Resource level

Hierarchical forecasting model
Forecasting Models

- Predicted variable is called the dependent or response variable

- Variables used to forecast the value are called the independent variables

- Causal models and trend models
  - Causal model: uses business indicators as the independent variables, workload parameters as the dependent variables
  - Trend model: Forecast a single workload variable as a function of time, using historical data and assuming the past trend will continue into the future
A business-oriented methodology

Main activities

- Select the Web service demands to be forecast
- Identify the business units associated with the services whose growth will be forecast
- Collect statistics concerning the chosen business units
- Summarize the statistics
- Translate business units into resource demands
- Forecast the future resource demand as a function of the business units
Forecasting Techniques

- Historical data can be identified as being either random, trend, or seasonal.

- For stationary data, the mean value is constant and the autocorrelation is only dependent on the relative position of the data (not time dependent).

- Idea is that information to be forecast is related to the historical data.
Regression Methods

Let $Y_1, Y_2, \ldots, Y_t$, be the time series that represent the observed values.

Develop a function to describe $Y_t$

$$Y_t = f(\bullet) + \epsilon_t$$

Causal model is represented by the following expression

$$Y = f(X_1, X_2, \ldots, X_n)$$
Forecast Error

- Forecast error for period $t$ is given by $\epsilon_t = Y_t - F_t$
- Mean error (ME) for $n$ periods of evaluation is:
  \[
  ME = \frac{\sum_{t=1}^{n} (Y_t - F_t)}{n}
  \]
- the mean of the squared errors (MSE) is:
  \[
  MSE = \frac{\sum_{t=1}^{n} (Y_t - F_t)^2}{n}
  \]
- the sum of the squared errors (SSE) is
  \[
  SSE = \sum_{t=1}^{n} \epsilon_t^2 = \sum_{t=1}^{n} (Y_t - F_t)^2
  \]
Linear Regression

General equation is:

\[ Y = a + bX \]

where \( Y \) is the dependent variable, \( X \) the independent variable, \( a \) the y-intercept and \( b \) is the slope.
Linear Regression

The regression parameters that minimize

\[ \sum_{t=1}^{n} (Y_t - F_t)^2 = \sum_{t=1}^{n} [Y_t - (a + bX_t)]^2 \]

are

\[ b = \frac{n \sum_{t=1}^{n} X_t Y_t - (\sum_{t=1}^{n} X_t)(\sum_{t=1}^{n} Y_t)}{n \sum_{t=1}^{n} X_t^2 - (\sum_{t=1}^{n} X_t)^2} \]

\[ a = \bar{Y} - b \bar{X} \]

where \((X_t, Y_t)\), are the coordinates of the \(n\) observed data points, \(\bar{Y} = (\sum_{t=1}^{n} Y_t) / n\) and \(\bar{X} = (\sum_{t=1}^{n} X_t) / n\)
Coefficient of determination

- Value is between 0 and 1 and gives an indication of how close the estimated $Y$ is to actual $Y$ values

$$R^2 = \frac{\sum_{t=1}^{n} (F_t - \bar{Y})^2}{\sum_t = 1^n (Y_t - \bar{Y})^2}$$
Examples

See page 467
Nonlinear methods

- Exponential method is useful when there is, or has been, increasing growth or decline of the workload parameter in past periods.

- Let

\[ Y_t = a \times b^t \]
\[ \log Y_t = \log a + t \log b \]

Let \( \log Y_t = Y' \), \( \log a = a' \), and \( \log b = b' \)

- Expression becomes

\[ Y' = a' + b' t \]

- Linear regression is used to solve \( a' \), and \( b' \)
Example

A financial Web service is growing in popularity. Table 12.5 shows the service request traffic for the peak second of the last ten months. Management wants to forecast future demand. Since Web traffic growth is assumed to be exponential, the following expression is used. $Y_t = a \times e^{bt}$

$\ln Y_t = \ln a + bt$. So let $\ln Y_t = Y'$ and $\ln a = a'$ so

$$Y' = a' + bt$$

Use linear regression to get $a' = 6.717$ and $b = 0.110$. However, $a = e^{a'} = 826.33$ so

$$Y_t = 826.33 \times e^{0.110t}$$
Simple moving average (SMA) is defined as

\[ F_{t+1} = \frac{Y_t + Y_{t-1} + \cdots + Y_{t-n+1}}{n} \]

where \( F_{t+1} \) is the forecast value at time \( t + 1 \), \( Y_t \) is the observation at time \( t \) and \( n \) is the number of observations used to calculate \( F_{t+1} \).

Example on page 473.
Exponential Smoothing

Similar to moving average, but places more weight on most recent observations

\[ F_{t+1} = F_t + \alpha(Y_t - F_t) \]

where \( \alpha \) is a smoothing weight from \( 0 \leq \alpha \leq 1 \) most typically \( 0.05 \leq \alpha \leq 0.3 \)
Summary

Basic steps of workload forecasting are:
- Definition of forecasting strategy
- Selection of the workload to be forecast
- Analysis of historical data and estimation of workload growth
- Selection of forecasting techniques
- Use of the forecasting techniques on historical data
- Analysis and validation of forecast results