Time Series Data Mining: 
A Retail Application

Daniel Hebert, Market Analyst, Rogers Corporation, Woodstock, CT, USA
Billie Anderson, Department of Mathematics, Bryant University, Smithfield, RI, USA
Alan Olinsky, Department of Mathematics, Bryant University, Smithfield, RI, USA
J. Michael Hardin, Dean, Culverhouse College of Commerce and Business Administration, University of Alabama, Tuscaloosa, AL, USA

ABSTRACT

Modern technologies have allowed for the amassment of data at a rate never encountered before. Organizations are now able to routinely collect and process massive volumes of data. A plethora of regularly collected information can be ordered using an appropriate time interval. The data would thus be developed into a time series. Time series data mining methodology identifies commonalities between sets of time-ordered data. Time series data mining detects similar time series using a technique known as dynamic time warping (DTW). This research provides a practical application of time series data mining. A real-world data set was provided to the authors by dunnhumby. A time series data mining analysis is performed using retail grocery store chain data and results are provided.

Keywords: Data Mining, Dunnhumby, Dynamic Time Warping, Time Series, Time Series Data Mining

1. INTRODUCTION

Data is being collected by businesses at a rate never encountered before through web sources, cellular phones and social media. The growth of internet businesses has led to a whole new scale of data processing challenges. Companies like Google, Facebook, Yahoo, Twitter, and Amazon now routinely collect and process hundreds to thousands of terabytes of data on a daily basis. The rise of the availability of retail data represents a significant change in the volume of data which can be processed.

In retail environments information regarding sales transactions is often recorded routinely. Such purchase transactions exist as observations within a data set. These records maintain a data and time associated with each transaction. Such time-ordered data presents new challenges for businesses and companies. Sequential data that is collected over a period of time is referred to as time series. When faced with large data sets that contain time series, organizations must employ new and sophisticated techniques for analysis. Traditional data mining methods and statistical techniques are often inappropriate when analyzing data that possess a time factor. This dilemma has led to the development and

DOI: 10.4018/ijban.2014100104
The rising importance of time series data mining techniques. In order to address the increasing prevalence of time-ordered data, researchers and analysts have begun seeking new analysis methods that account for information collected over a period of time. In order to support informed decision making and gain useful knowledge, organizations are searching for novel means of understanding and interpreting time series.

Time series analysis consists of methods for analyzing time series data to extract meaningful information. Time series data mining combines data mining techniques with time series analysis to:

- Conduct similarity analysis of time series data to validate the forecasting of new products.
- Gain a deeper understanding of key markets of interest and notable buying behaviors.
- Identify products that share similar purchase rates.
- Improve predictive modeling capabilities.
- Achieve better customer retention and satisfaction.
- Enhance goods consumption ratio.

Time series is traditionally concerned with identifying trends in the data such as seasonality or trends and forecasting. Data mining is the process of detecting hidden relationships and patterns in very large data sets. Data mining focuses on developing predictive models and using nonparametric techniques such as clustering. The field of data mining includes methods that attempt to automate the scientific discovery process. The unique characteristic of data mining is the types of problems faced—those with large data sets that contain complex and hidden relationships.

The difference between time series and applying data mining techniques to time series data is the sheer amount of data that is involved. In time series data mining the number of time series that are available to analyze is so large that traditional time series methods are not feasible (Liu, Bhattacharyya, Sclove, Chen, & Lattyak, 2011).

The quantity of the time series is not the only reason that traditional time series methods are not applicable in certain situations. Traditional time series methods such as Box-Jenkins and the Autoregressive Integrated Moving Average (ARIMA) model both assume the time series are stationary; that is the time series remains constant through time (Box & Jenkings, 1976; Pandit & Wu, 1983). For real-world time series such as customers purchasing items or stock market movements, stationarity is not a realistic assumption. The ARIMA model also assumes that the system generating the time series is linear; that is the system can be described by difference equations (Gabel & Roberts, 1980). Many times in practice, the system generating the time series is not linear.

Time series data mining is a contribution to both time series and data mining disciplines. Time series data mining is able to overcome the limitations of traditional time series analysis such as linearity and stationarity by adapting data mining methods (Povinelli & Feng, 1999). The framework for which time series data mining is set allows the methodology to innovate data mining methods by allowing time series to be predicted, similarities quantified, and irregular time series to be handled more effectively than in the past. This framework allows applications such as predicting when the customers will buy next, detecting similar purchasing patterns over time for certain products, or determining when the customer will churn.

This article will examine how to use a time series data mining analysis to detect similar time series using real-world grocery retail data. The authors provide background related to time series data mining and specifically discuss how the retail industry are using these techniques to solve real business problems. A description of the algorithm that will be used to detect similar time series will be provided. The authors provide a description, analysis, and results of a real-world retail data set in which a time series data mining technique is applied in order to detect similar time series. The paper concludes with suggestions for future time series data mining applications.
Data Mining Applications in Accounting and Finance
Wikil Kwak, Susan Eldridge and Yong Shi (2014). Encyclopedia of Business Analytics and Optimization (pp. 609-617).
www.igi-global.com/chapter/data-mining-applications-in-accounting-and-finance/107264?camid=4v1a