

Mining coherence in time series data

George Potamias^{1,3} and Vassilis S. Moustakis^{1,2}
{moustaki, potamias}@ics.forth.gr

¹ Institute of Computer Science, Foundation for Research and Technology – Hellas (FORTH), Vassilika Vouton, P.O. Box 1385, 71110 Heraklion, Greece

² Department of Production and Management Engineering, Technical University of Crete, Chania 73100, Greece

³ Department of Computer Science, University of Crete, Heraklion 71409, Greece.

Abstract

This paper presents work on modeling coherence between time series data. Work is based on the elaboration of formula that computes the distance between time series. Based on the computed distances, the method exploits the closest neighbor algorithm and leads to the construction of a phylogeny-clustering tree. Using car sales data (available on the www) we demonstrate work done and present preliminary results. We discuss implications of the work performed in correlating time series data with documents including such data.

Introduction

Time series occur in many aspects of economic and social activity. Time series organization of data implies time stamping of individual observations. Observations may represent economic or social activity or even life critical measurements such as those acquired when a patient is under monitoring in a critical care unit. Time series data modeling has been an active area of research in statistics. A variety of models exist, which manifest interest and provide the interested analyst with analysis tools (Box and Jenkins 1976). Expanding interest in data mining and knowledge discovery has contributed to an increase of research awareness in learning using time series data (Morik 2000).

This paper reports preliminary research results and modeling activity in integrating time series data with documents. Motivation originates from the fact that analysts do use time series data to prepare reports. Reports and time series data reside in distributed information archives. There is an emerging need to mine for knowledge over such archives, which contain both reports and data. Specifically, the paper presents a methodology, which investigates similarity between time series data and reports results using exemplar time series drawn from car sales (see legend in Figure 1, in the text). Data-mining methodology is realized by the introduction of a novel algorithmic process and related formulas, for discovering similar and indicative patterns in time-series collections. Final outcome is the clustering of time-series into similar-groups, visualized by the appropriate customization of a phylogeny-based clustering algorithm and tool.

Methodology

Measuring similarity between objects is a crucial issue in many data retrieval and data mining applications. The typical task is to define a function $dist(a,b)$ (or, $sim(a,b)$), between two sequences a and b , which represents how “similar” they are to each other. For complex objects, designing such functions, and algorithms to compute them, is by no means trivial. Time series are an important class of complex data objects, and the need for clustering sequential data is profound. While the statistical literature on time series is vast, it has not studied *similarity* notions that would be appropriate for data mining applications.

Our *methodology*, for discovering coherences between time-series is realized by the following steps:

- i. *Representation*. A *Piecewise Linear Segmentation* approach is followed. The approach aims to ‘weight’ the different time-to-time changes (segments) of time-series, based on their ‘significance’ according to the whole

evolution of the series, i.e. the *intra-significance* of a change. Given a time-series X_a , evolving over a set of k time-points with respective values y_1, y_2, \dots, y_k , the *percentage change*, $c_{a,i;i+1}$ is computed by the formula:

$$c_{a,i;i+1} = \frac{|y_{i+1} - y_i|}{y_i} \quad (1)$$

The significance of a change is decided by the applying a suitable statistical-test. In the current implementation, and in the conducted experiments the *Student's t*-distribution and respective statistics were used.

- ii. *Weighting percentage changes.* Besides measuring the intra-significance of changes in a time-series, the respective changes should also be weighted according to the respective time-to-point changes for the whole time-series collection, i.e., the *inter-significance* of a change. The *relative-weight* of a time-series change is computed by the formula,

$$w_{c_{a,i;i+1}} = \frac{c_{a,i;i+1}}{\sum_{\substack{\text{Time series} \\ \text{in } \{a, b, \dots\}}} c_{T,i;i+1}}, \quad (2)$$

- iii. *Computing the distance between time-series.* Given two time-series X_a and X_b , their distance is computed by the formula:

$$\text{dist}(X_a, X_b) = \sum_{i=1}^{k-1} (w_{c_{a,i;i+1}} c_{a,i;i+1}) + (w_{c_{b,i;i+1}} c_{b,i;i+1}) \quad (3)$$

Utilization of inter-significance. If the respective percentage changes are significant then, their weights are computed by formula 2 above. In all other cases, the weights are not taken in consideration, i.e., their values are fixed to 1.

Clustering and Visualization. The computed distances feed an appropriate distance-based clustering algorithm in order to form clusters of similar time-series. The current implementation of our system encompasses the Neighbor-Joining clustering algorithm (Saitou and Nei 1987). The *Neighbor-Joining* algorithm is a phylogeny-based clustering algorithm utilized in biology-systematics and molecular biology. The main utility of the algorithm is that it computes, and outputs a phylogeny-tree that offers a nice and informative *visualization* of the formed clusters.

Experimentation

Six time series data depicting car sales were randomly selected. Data are indicated as TS-1 through TS-6 and are presented in Figure 1. In each time series the horizontal axis presents time periods and the vertical axis sales.

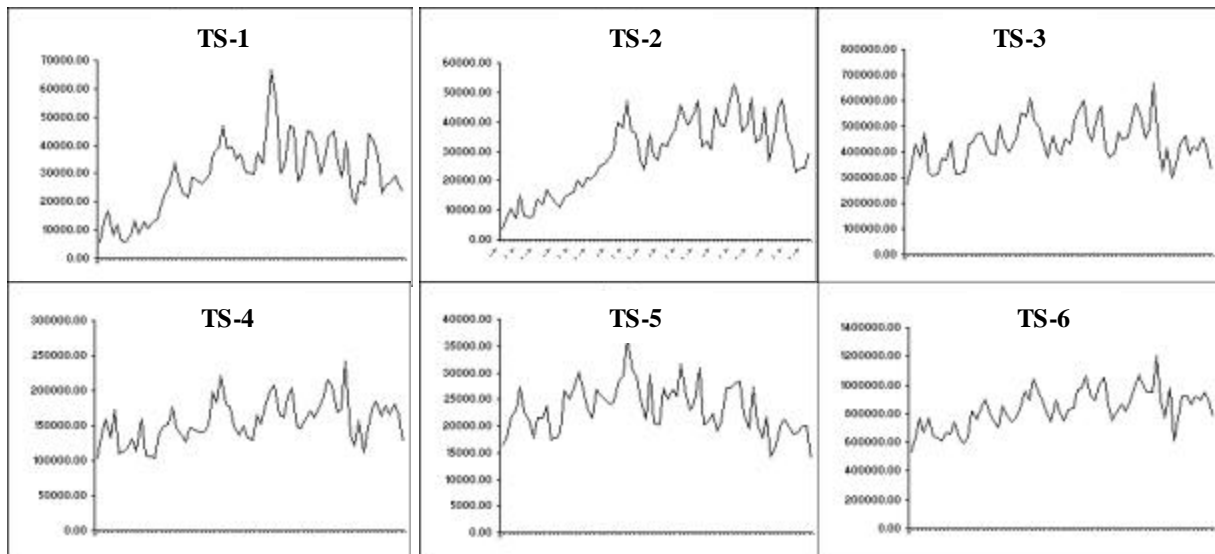


Figure 1: Car sales time series data (M2-competition, <http://forecasting.cwru.edu/Data/m2comp>)

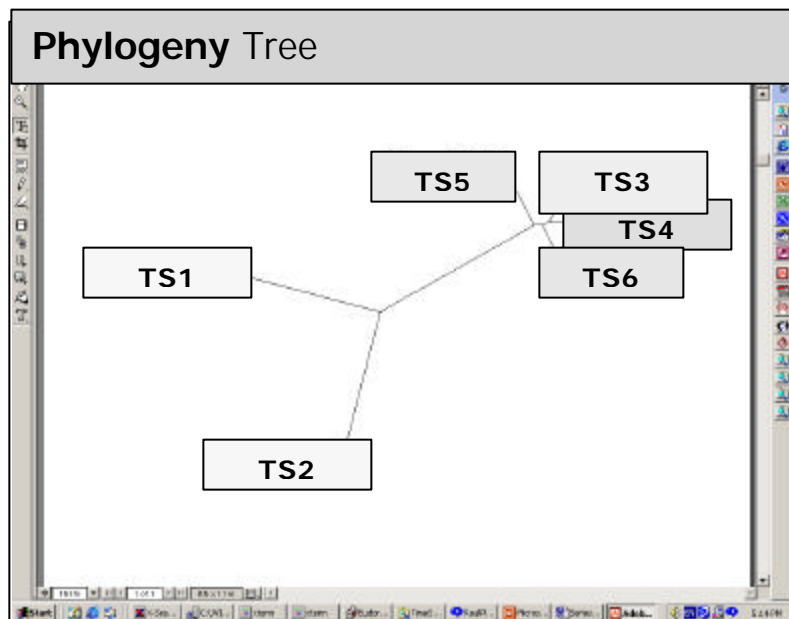


Figure 2: *Phylogeny tree* of the car sales data (see Figure 1). Time series 1 and 2 (*Cheetah and Panther*) are grouped together and isolated by all other time-series. Time series 3 and 4 (*Animal and Bigcat*) are considered similar and grouped together. Time-series 5 and 6 (*Lion and Carrinds*) presents isolated not-coherent objects. Furthermore, time-series 3,4,5, and 6 are isolated from 1 and 2 and could be considered as a sub-group in the hierarchy of clusters.

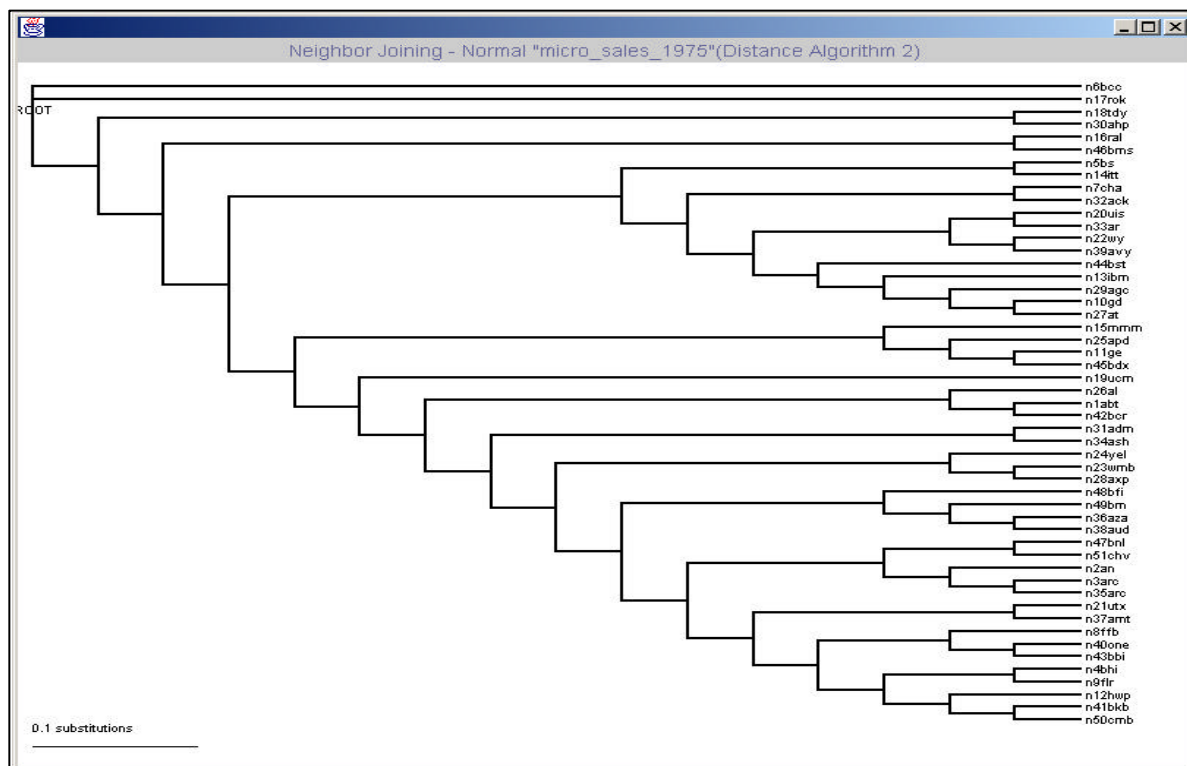


Figure 3: *Dendrogram* visualization of phylogeny-based clustering of multiple time-series

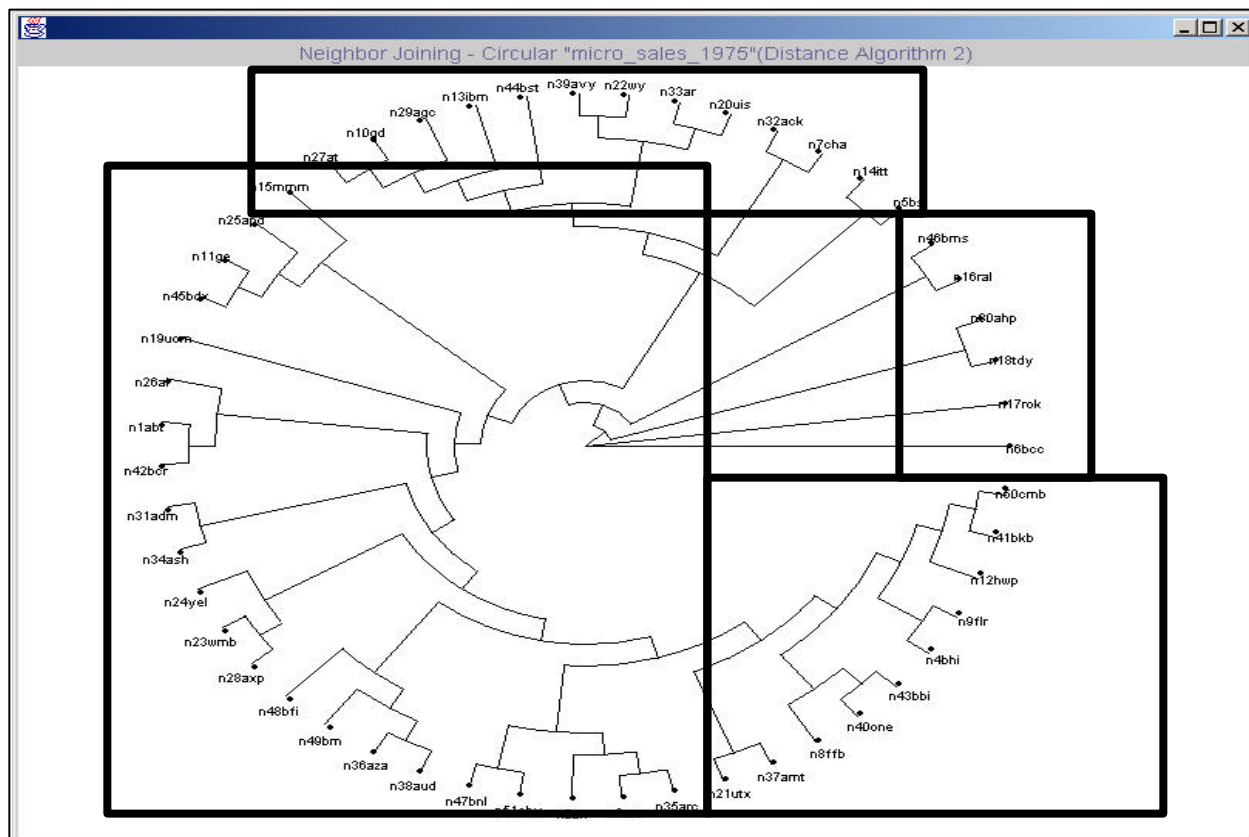


Figure 4: Circular visualization of phylogeny-based clustering of multiple time-series – Identified clusters

Application of the algorithm, which we present in section on Methodology led to the construction of the phylogeny-clustering tree that we depict in Figure 2. Broader application of the methodology on a wider collection of time series data led to the construction of the phylogeny-clustering tree, which we present in Figure 3. Figure 4 presents an alternative visualization of the phylogeny tree construction. Results point out the feasibility and potential of the selected approach.

Discussion

Preliminary results reported in this paper demonstrate feasibility of an alternative modeling approach to identify similarity and coherence across time series data. Usefulness of the approach in identifying the most relevant time series and corresponding documents (if any) across large (and often distributed) statistical collections is critical. The user tries to identify both data and documents. Modeling coherence across time series, which have been used to prepare documents (such as the documents reporting on industrial or economic activity in a region or across industries) eases retrieval and contributes to the location of the most appropriate information. In other words, modeling coherence provides an alternative indexing approach to a huge collection of both time series and documents.

We are applying this approach in context of a European project titled IRAIA (IRAIA 2001) -- IRAIA stands for *Getting Orientation in Complex Information Spaces as an Emergent Behavior of Autonomous Agents*. Task is to support users to access documents and time series data across repositories that contain thousands of both.

The implemented system offers operations for importing and reading time-series data; computing distances between time-series; clustering of time-series; and visualization of the resulted clustering. The whole system is built in *Java*

making it *portable* on various h/w platforms (the java classes of the clustering algorithm were taken from the public-domain library PAL- Phylogeny Analysis Library, <http://www.pal-project.org>, and were appropriately customized for inclusion in our system).

Furthermore, the approach may be extended to modeling cause-effect relationships between variables, each of which is represented via a different time series.

We maintain that modeling and learning from time series data represents an emerging machine learning domain, which is further motivated by data mining inquiries.

Acknowledgement

Work presented in this paper is partially supported by the European Union, IST Program via the IRAIA project (IST 1999 10602). The partners of the IRAIA project are: DIW (German Institute for Economic Research, D), IFO (Institut für Wirtschaftsforschung, D), CNRS-UPS (Centre National de la Recherche Scientifique, F), QUB (Queens University of Belfast, UK) and FORTH (Foundation for Research and Technology – Hellas, GR). Responsibility for work presented lies with the authors and not with the IRAIA consortium.

References

Box, G.E.P. and Jenkins G.M. 1976. *Time Series Analysis, Forecasting and Control*. Prentice Hall

IRAlA. Getting Orientation in Complex Information Spaces as an Emergent Behavior of Autonomous Agents. Overview information available at: [www](http://www.iraila.org).

Morik K. 2000. The Representation Race – Preprocessing for Handling Time Phenomena. In: Mantaras, R.d.L. and Plaza, E., *Proceedings of European Conference on Machine Learning (ECML)*. Springer Verlag

Saitou, N. and Nei, M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4** (4), 406–425