Temporal-PEx: Similarity-based Visualization of Time Series

A. B. Alencar, Maria C. F. de Oliveira, F. V. Paulovich, R. Minghim, and M. G. Andrade
University of São Paulo
Institute of Mathematics and Computer Science
{aretha,cristina,paulovic,rminghim,marinho}@icmc.usp.br

Abstract

Time series analysis poses many challenges to professionals in a wide range of domains. Several visualization solutions integrated with mining algorithms have been proposed for exploratory tasks on time series collections. As the data sets grow large, though, the visual alternatives do not allow for a good association between similar time series. In this paper we introduce a visual representation or large time series data sets generated by multidimensional projections based on distance measures.

1. Introduction

Time series occur extensively in the real world, in application domains as diverse as epidemiology, biostatistics, and engineering. Huge time series databases become widely available as computing becomes ubiquitous, and their analysis has deserved increased attention. Visualization has been successfully used to analyze time series data for a long time. However, time series are a difficult type of data to visualize because of their natural high dimensionality. In time series mining, the term 'dimensionality' indicates the series size. This high dimensionality associated with huge datasets favors visual occlusion.

Projection techniques are a fine example of supporting tool for visual data mining of high dimensional datasets. The goal of projection techniques is to create a two-dimensional display from data originally represented in higher dimensional spaces, while striving to lay out largely similar points in a way that they can be visually recognized as highly related. These visual representations allow users to employ their visual ability to interact and explore the data in search of patterns, trends, or outliers. Point positioning is determined by the multidimensional projection technique, according to a time series dissimilarity measure.

Projections are the major visualization resource in a tool called *Projection Explorer* (PEx), which provided the basis for this work. PEx was also originally designed to examine text collections [5], but encompasses projections that are suitable for any space that can be embedded in a metric space. We extended PEx to deal with detailed information of time series, creating a projection based time series analysis tool – *Temporal Projection Exmplorer* (Temporal-PEx). This tool supports the user to have a general view of behavior of time series collections, as well to find groups of series with similar behavior.

2. Our tool: Temporal Projection Explorer

A projection wizard is provided to guide a user in creating a new projection. The first step is to define a time series data source, thas an organization similar to the *Comma Separated Values* (CSV) file format. Some type of data preprocessing is typically required to improve projection accuracy, and this is the next step handled by the wizard. Three pre-processing procedures are available: a time series normalization procedure, a noise reduction procedure, and a procedure to remove seasonal patterns from the time series.

The third and final step in creating a projection is to choose a combination of projection technique and time series dissimilarity measure. Two projection techniques are embedded in Temporal-PEx: IDMAP (*Interactive Document Map*) [4] and LSP (*Least-Square Projection*) [6]. IDMAP is the default and more precise technique, but it is $O(n^2)$. LSP is a precise and fast technique, but it favors the formation of clusters, which is not always a desirable thing. And there are three options for time series dissimilarity measures: the *Euclidean* distance, DTW (*Dynamic Time-Warping*) [1], and the CDM (*Compression-based Dissimilarity Measure*) [3]. The Euclidean distance is a rapid and common distance, but it can only address time series with same size. For time series with different sizes or distortions in the time axis, the DTW distance is more indicated. The CDM distance is used to detect structural differences, and only works in long time series.

The resulting projection is shown at the main window panel of the tool (see Figure 1). The main window includes three panels: the projection panel (center) shows the projected time series, where each point corresponds to a single time series; a time series labels panel (top left corner) shows descriptive labels associated to each time series; and a details-on-demand panel (bottom left corner), that it is a tabbed panel that displays tree types of information about a time series: its data values, its neighboring points, i.e., the series most similar to it in the data, and statistics.

At the top toolbar two combo boxes allows mapping one of five time series attributes to the color and size of the points: mean value, variance, minimum value, maximum value or class. Assuming that a classification of the time series collection is known, the class attribute may be supplied in the time series data file. A double click over a particular point will open a window that displays the line chart of its corresponding time series.

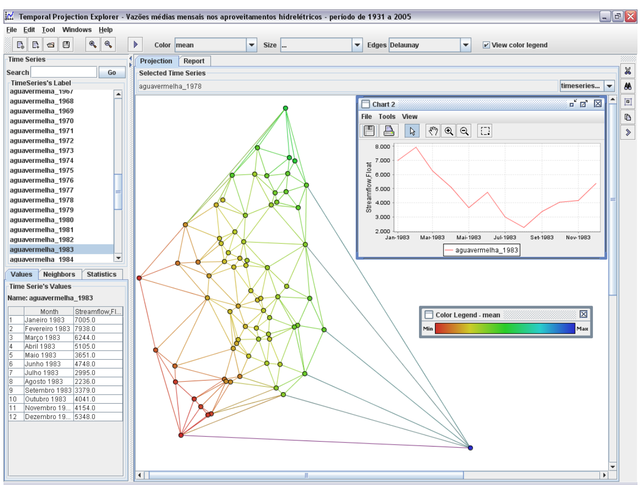

Figure 1. Screenshot of Temporal-PEx.

3. Application example: streamflows of brazilian hydroelectric power plants

The Brazilian electric power distribution system is supplied mainly by hydroelectric power generation, which is heavily affected by seasonal rain patterns. Analysis of streamflows of different hydrographic basins collected along the years plays a strategic role in planning and operation of the national electric distribution system. This data set contains real time series describing the natural streamflow of hydroelectric power plants installed in the main hydrographic basins in Brazil [2]. For each power plant a time series is available that registers the monthly average streamflow values for a period of 75 years, from 1931 to 2005. Some of the data values are estimated. Figure 2 depicts the projection of the 76 power plants in the dataset from the Paraná basin. These time series were normalized and had their seasonality removed in the pre-processing step. It is possible to clearly identify six regions that clearly correspond to the five sub-basins, labeled by different colors.

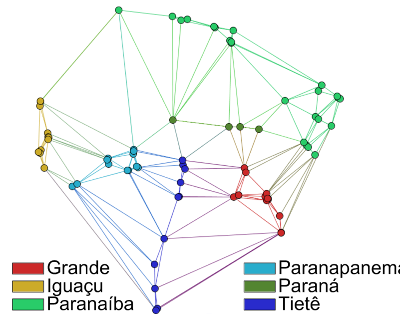

Figure 2. Power plants of the basin Paraná

4. Conclusions and future work

In this paper we contribute by demonstrating the usefulness of projection techniques in identifying patterns in temporal data such as large time series describing power plants streamflows. We plan to incorporate addition time series mining algorithms, such as classification and discovery of motifs – previously unknown frequent patterns.

References

[1] D. J. Berndt and J. Clifford. Using dynamic time warping to find patterns in time series. In *Workshop on Knowledge Discovery in Databases*, 1994.

[2] O. N. do Sistema Elétrico. Vazões médias mensais nos aproveitamentos hidrelétricos - período 1931 a 2005. Technical report, 2007.

[3] E. Keogh, S. Lonardi, C. A. Ratanamahatana, L. Wei, S.-H. Lee, and J. Handley. Compression-based data mining of sequential data. *Data Mining and Knowledge Discovery*, 14:99–129, 2007.

[4] R. Minghim, F. V. Paulovich, and A. A. Lopes. Content-based text mapping using multi-dimensional projections for exploration of document collections. In *IST/SPIE Workshop on Visualization and Data Analysis*, 2006.

[5] F. V. Paulovich and R. Minghim. Text map explorer: a tool to create and explore document maps. In *Proceedings of the Information Visualization Conference*, 2006.

[6] F. V. Paulovich, L. G. Nonato, R. Minghim, and H. Levkowitz. Visual mapping of text collections through a fast high precision projection technique. In *Proceedings of the conference on Information Visualization*, 2006.