SIMILARITY SEARCH USING MRDTW (MAP REDUCE DYNAMIC TIME WARping) IN HADOOP
S.Naveen¹, M. Pajany²

¹Research Scholar, Department of Computer Science Engineering IFET College of Engineering Villupuram, India
²Associate Professor, Department of Computer Science Engineering IFET College of Engineering Villupuram, India
naveenrajmass@gmail.com
pajanyanish@gmail.com

Abstract—The proposed the MRDTW parallel algorithm which is able to search two large time series more accurately. The MRDTW algorithm works by breaking a larger time series into sub sequences and then compares their relationship in the segmented time series. The sub sequences can be compared independently which allows us to use parallel schemes like map reduce. This paper aims to improve the performance of the DTW algorithm using parallelism. Experimental evaluation results indicate that the technique retains the searching accuracy of the DTW with a much improved similarity measure. Using Hadoop clusters parallel processing was the following benefits. We can improve the speed of the analysis and we can scale the cluster by adding additional cluster nodes. Other benefits are cost efficiency and high availability and resiliency to failure. When a piece of data is sent to a node for analysis, the data is also replicated to other cluster nodes. That way, if a node fails, additional copies of the node’s data exist elsewhere in the cluster and the data can still be analyzed.

Keywords: MRDTW, DTW, Hadoop, Map Reduce.

I. INTRODUCTION

Time series is an important class of temporal data objects and it can be easily obtained from scientific and financial applications. A time series is a collection of observations made chronologically. The nature of time series data includes: large in data size, high dimensionality and necessary to update continuously. The increasing use of time series data has initiated a great deal of research and development attempts in the field of data mining. Euclidean Distance is very fragile to small temporal distortions and, in some particular application, notch distance measure is required. Dynamic Time Warping (DTW) improves those weaknesses by reaching optimal alignment between multiple time series, which are warped in a nonlinear fashion to match with each other. Therefore DTW can be used to measure the similarity in time series with patterns shifted in time or distorted in sharp. In this paper we introduce a technique which speeds up DTW by a large constant. The value of the constant is data dependent but is typically one to three orders of magnitude. The algorithm, Segmented Dynamic Time Warping (SDTW), takes advantage of the fact that we can efficiently approximate most time series by a set of piecewise linear segments. The main process of the paper that degrading the performance is Time and space complexity, it is hard for DTW to measure similarity in large time series with more than thousands of data points. The growing volume of data sets make the implementation of mining algorithm complicated in modern medical and industrial database. It is not be sufficient for scaling DTW methods to truly massive databases.

II. RELATED WORKS

In this section, we have studied previous research papers related to the similarity search using MRDTW. The related works of this study. The variety of approaches have been suggested. These approaches are robust to noise, offset translation and amplitude scaling to varying degrees. However, they are all extremely sensitive to scaling in the time axis (longitudinal scaling). We present a method for similarity search that is robust to scaling in the time axis, in addition to noise, offset translation, and amplitude scaling. The method has been tested on medical, financial, space telemetry and artificial data. Furthermore the method is exceptionally fast, with the predicted 2 to 4 orders of magnitude speedup actually observed. The method uses a piecewise linear representation of the original data. We also introduce a new algorithm which both decides the optimal number of linear segments to use, and produces the actual linear representation.

Visualizing and Discovering Non-Trivial Patterns In Large Time Series Databases

AUTHORS: Michael J. Pazzani , University of California at Irvine
Data visualization techniques are very important for data analysis. The visually summarizes both the global and local structures of time series data at the same time. The problem due to summarizing the data it has a technician making i.e. go/no-go decision. The possibility of finding go/no-go decisions consumes high time.

A indexing scheme for fast similarity search in large time series database

AUTHORS: Eamonn J. Keogh, University of California at Irvine

The index is formed by creating bins that contain time series subsequences of approximately the same shape. The bound allows us to search the bins in best first order. DTW algorithm has a quadratic time and space complexity.

A Simple Dimensionality Reduction Technique for Fast Similarity Search in Large Time Series Databases

AUTHORS: Eamonn J. Keogh and Michael J. Pazzani

The dimensionality reduction technique that supports an indexing algorithm that is more than an order magnitude faster than the previous best known method. The ability to interpret the influence of individual features/variables goes down while applying PCA. It cannot identify an exact attribute or dimension in the component.

III. PROPOSED SYSTEM

In this paper, we proposed to efficiently perform fast-similarity search in parallel using the popular MapReduce framework. We proposed the Map-Reduce-based dynamic time warping (MRDTW) parallel algorithm is developed, which is able to search through a more accurate warp path between two large time series. The working principle of MRDTW algorithm is achieved by breaking the larger series into sub-sequences and comparing their similarities in the segmented time series. The comparison of sub-sequence is independent, which allows the MRDTW to use a parallel scheme like MapReduce. Unlike the previous methods modified the DTW algorithm, the main contribution of this paper is trying to improve the performance of DTW using the parallel scheme when dealing with large time series. Experimental evaluation results illustrate our approach retains the searching accuracy as the DTW, and the efficiency of similarity measure has been greatly improved in large time series. MRDTW(Map-reduce dynamic time warping) the parallel algorithm search two large series more accurately. The subsequence can be compared independently. Similarity search used to cluster, execute data process bound.

IV. SYSTEM IMPLEMENTATION

Module 1: Hadoop Configure

In this Hadoop module first assume gk-environment variable and the assume the hadoop environment variable in dash file. Then extend the hadoop in all the nodes. Configure the HOFT in all the nodes and also configure the SLAVE in namenodes.

Module 2: HDF Implementation

This module describe about the HDF. Format the name node using HDF first start, yarn-DAEMON file present in the HDF and also yarn-DAEMON file in the name node.
Module 3: Run Mapper and Reducer Drive File

This module describes about the output format.

VI. RESULTS

We demonstrated a modification of DTW that exploits a higher level representation of time series data to produce one to three orders of magnitude speed-up with no appreciable decrease in accuracy. We experimentally demonstrated our approach on several real world datasets. Future work includes a detailed theoretical examination of SDTW, and extensions to multivariate time series.

REFERENCES


