A Similarity-Based Method for Visual Search in Time Series Using Coulomb's Law*

Claudinei Garcia de Andrade and Marcela Xavier Ribeiro

Universidade Federal de São Carlos,
São Carlos-SP, Brazil
{claudinei.andrade,marcela}@dc.ufscar.br

Abstract. We present a method for visual search in multidimensional time series based on Coulomb's law. The proposed method integrates: a descriptor based on Coulomb's law for dimensionality reduction in time series; a system to perform similarity searching in time series; and, a module for the visualization of results. Experiments were performed using real data, indicating that the proposed method broadens the quality of through similarity queries in time series.

Keywords: Time series analysis, Index method, Similarity Search, Coulomb's law.

1 Introduction

The great early challenges to work with the analysis of temporal observations is the development of compact storage methods for series that are truly representative of the collected information, which are easy to handle and show a high level of accuracy for knowledge extraction. Thus, this paper aims to propose an integrated environment for similarity search in time series with the incorporation of a descriptor based on Coulomb's law for dimensionality reduction. In addition to it, the paper presents a system to perform similarity searching in time series and also a module for the visualization of results. Experiments with real data of varying sizes and dimensions provide validation and confirm that the system produces satisfactory results.

2 Background and Related Works

A time series can be defined as an set of observations [1], $\{Y(t), t \epsilon T\}$ in which Y is the variable of interest and T is an index set. Time series are considered complex data. There are not any way of establishing an order relation among series or their ranges. In this context, the concept of similarity is more applicable than the concept of equality.

Current similarity search methods are, in general, based on the use of descriptors in order to obtain similarity ranges. Some authors define a descriptor

* We would like to thank FAPESP, CAPES and CNPq for the financial support.

A.J. Machado Traina et al. (Eds.): SISAP 2014, LNCS 8821, pp. 241–246, 2014.
DOI: 10.1007/978-3-319-11988-5_22

as being formed by a pair (ϵ_D, δ_D) where ϵ_D is the component responsible for characterizing the object through the extraction of characteristics and generating a vector that will be used to analyze the data. δ_D is the function responsible for comparing the characteristics vectors, giving the amount of similarity between the object and the query [2]. There are a variety of descriptors that are effective for certain data fields but end up presenting loss of representativeness of the series data in most cases [3].

There are two basic types of similarity queries: i) (*Range queries*) which finds objects that are at a maximum r distance of the object query Q; and ii) (*k-Nearest Neighbor query* or *k-NN query*) which aim to retrieve the k objects most similar to a query object.

3 Proposed Method

The proposed system is composed of distinct modules that share data with one another and work harmoniously getting the information passed by the user to carry out the queries, applying the Coulomb descriptor to the data according to the user's interest and graphically returning objects of interest as found and listed by the descriptor.

Figure 1 illustrates the relationships among modules. Time series data serve as input to the visualization and data exploration module (VDEM) where the expert can verify the behavior and relevant characteristics of the series and select the interesting intervals for analysis. Also, they serve as input to the Coulomb descriptor module (CDM) which, by dimensionality reduction and similarity calculation, passes the ranges with some degree of similarity on to the data analysis module (DAM), according to the user's interests. From there, the data analysis module prepares information that is handed back to the VEDM, which, in turn, displays them to the user.

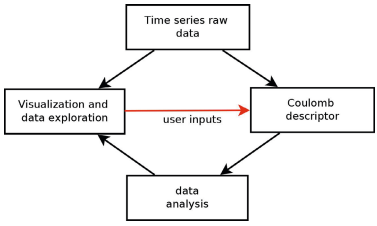

Fig. 1. Relationships among modules

Coulomb's law establishes the mathematical relationship between the charges of two or more bodies and their electrical force output by calculating the existing interaction forces (attraction and repulsion) in these charges. The principles of Coulomb's law can be expressed thus: i) the intensity of the electric force is directly proportional to the product of the electric charges; and ii) the intensity of the electric force is inversely proportional to the square of the distance between the bodies. The law's formula is: $\boldsymbol{F} = K\frac{q_1 q_2}{r^2}\hat{r}$. Where: $\boldsymbol{F}$ is the force in newtons; r is the distance between the two point charges; q_1 and q_2 are the intensities of the charges; $\hat{r}$ is the unit vector; and K is Coulomb's constant.

This proposal for similarity search in series considers the observations of the time series as point charges with q constant charge values. Those are located in

the coordinate plane formed by the series index and by the value of the observation. Since calculating the distance between the charges is necessary in order to ascertain the interaction between them, we take a Cartesian plane formed by the time series index (x-axis) and the value of the observations (y-axis). Thus, it is possible to calculate the distance between the charges for the calculation of forces.

Furthermore, a dummy point charge, of charge q, is inserted into the centroid composed of the sets of observations that make up the search ranges. This charge is aimed at providing an optimal representation of the range because, in addition to its being located at the geometric center of the range, it is used to calculate the interaction between itself and the other charges, thus generating the resultant force that represents the range.

As the resultant force is a vector measure, the charge's direction and magnitude influence the calculation. So, it was established that charges that are below the centroid charge are in the opposite direction to those above it, consequently, posing a negative force. Accordingly, it is possible to represent a time series through a system of electrically charged particles and to calculate the resultant force F, obtained through a vector sum of all forces that comprise the system. That way, we are able to reduce the series dimensionality, contributing to similarity search without major loss of information. In the proposed approach, the feature vector ($V = [\boldsymbol{F}, h]$) is formed both by the resultant force calculated in the range of interest and the height of the centroid.

4 Experimental Results

The experiments devised to test the proposed environment were divided into two groups: i) experiments with the Coulomb descriptor and DAM to validate the descriptor's performance in reducing data dimensionality and with DAM in finding windows with higher similarity in the series; ii) experiments with VDEM integrated with other modules. We used randomly generated databases, a meteorological database of several Brazilian cities with minimum and maximum temperatures, along monthly precipitation indexes from the years 1961 to 2010 obtained in [8], as well as medical data obtained in [9] regarding glucose levels in patients in the course of their daily activities.

	DFT	SM	Coulomb
Accuracy	20,48%	46,63%	68,95%

Fig. 2. Descriptors' accuracy comparison

4.1 CDM and DAM Tests

Validating the Coulomb descriptor and the data analysis module aims to verify the performance of the descriptor in reducing data dimensionality and in finding windows with higher similarity in the series. We saw fit that the Coulomb descriptor was compared to the Sequential Matching *Sequential Matching* (SM) [4], [5] and *Discrete Fourier Transform* (DFT) [6] methods, since those methods are

considered baselines of the work in question. The former for presenting high accuracy and the latter for having good performance in large databases. The modules were evaluated in the following aspects:

1) Computational Complexity: in this respect, we performed two experiment. The first, was a runtime test of the same *knn* query using three descriptors, with varying database sizes, graph in Figure 3. The second was a runtime test of a *knn* query with varying query window sizes, in Figure 4, we note that the Coulomb descriptor presents a shorter runtime as compared to the SM and DTF descriptors.

2) Accuracy: intended to measure the number of instances that were predicted correctly from an input query. In this test, we consulted the most similar periods (*knn*-query) to the period encompassing summer (December 21 to March 20 of the following year) and winter (from June 21 to September 21) in the city of Araraquara/SP. Queries were carried out using the descriptors mentioned and the accuracy results are presented in table of Figure 2. As shown, the accuracy displayed by the Coulomb descriptor is satisfactory for similarity queries.

3) Precision vs. Recall: proposed by [7]. From the meteorological base, we used data concerning monthly maximum temperatures of Presidente Prudente/SP, Brazil. The three aforementioned descriptors were implemented with a focus on similar seasons, periods of abnormal increase or decrease in temperatures and periods with some cyclical temperature variability. The graph in Figure 5 represents the precision and recall found. With the medical base, the tests searched periods of high and low blood glucose levels in patients before and after the administration of insulin as well as before and after meals. The precision vs. recall graph is shown in Figure 6. By analyzing the graphs, we note that the accuracy of the Coulomb descriptor remains satisfactory for good levels of recall, if compared to other methods.

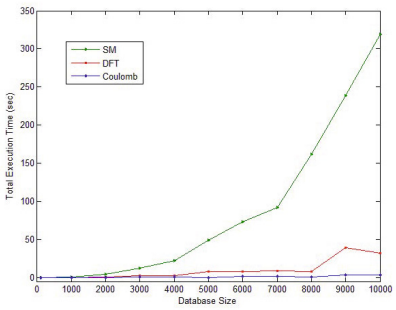

Fig. 3. Runtime per query with varying database sizes

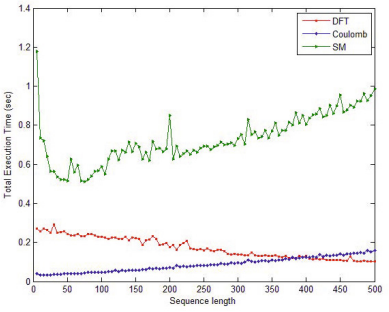

Fig. 4. Runtime per query with varying query window sizes

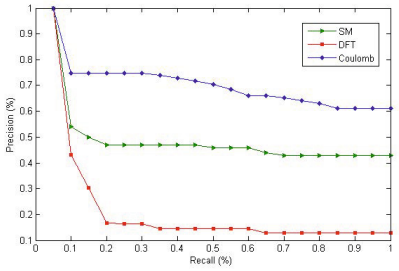

Fig. 5. Precision vs. Recall - meteorological database

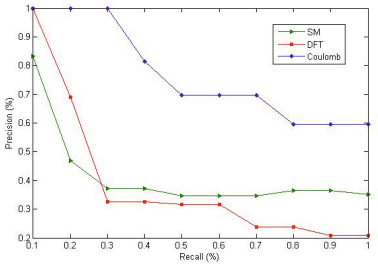

Fig. 6. Precision vs. Recall - medical database

4.2 Viewing Environment Test

For carrying out tests on the visualization and data exploration module, similarity queries in meteorological time series were run for winter and summer periods. The query shown in Figure 7 was run on the time series containing the average temperature of the city of Araraquara/SP between the years 1979 and 2010. A *knn*-query with $n = 10$ and the winter of 1979 as period of interest (leftmost hatched period in the graph). As can be seen in the figure, the periods returned (hatched portions of the graph) by the system correspond to winter periods where there was a minimum temperature close to the selected range.

Another test carried out uses three time series regarding the maximum monthly temperatures of the cities Avaré, São Paulo and Presidente Prudente in the years 1970-2008. The similarity query with $knn = 10$ is run by selecting the winter of 1988 in Presidente Prudente as the period of interest. As shown in Figure 8, periods of greatest similarity concerning the three series are hatched.

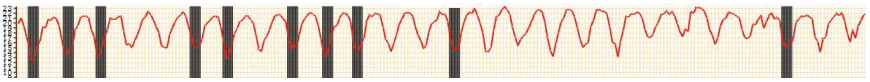

Fig. 7. knn query = 10 applied to winters in the city of Araraquara/SP

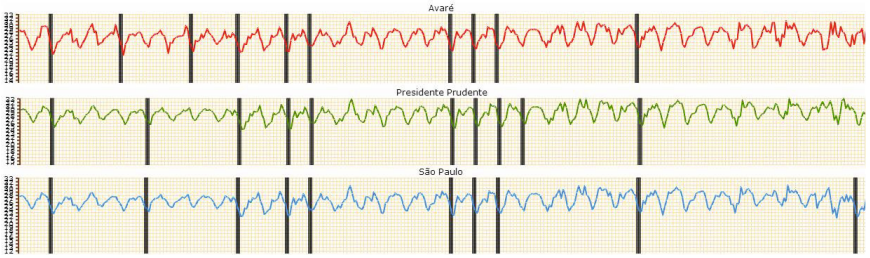

Fig. 8. knn query = 10 applied to the winter of 1988 in Presidente Prudente/SP

As shown, the visual module provides satisfactory results and allows a specialist to view similar ranges in an understandable and practical way. It makes it possible for similarity queries to be used in inferring knowledge about the series under analysis.

5 Conclusions

Upon analysis of the results obtained, we reach the conclusion that the Coulomb descriptor presents satisfactory values of accuracy and runtime for the execution of similarity queries on time series. Furthermore, a comparison of the Coulomb descriptor with traditional search methods for time series, through the analysis of accuracy vs. recall graphs, reveals significant advantages. That makes the Coulomb descriptor a potential descriptor for time series in different areas. In addition, the visualization and data exploration module allows a specialist to perform similarity queries. As a future task, the current modules will be integrated into a data-mining module in order to generate association rules using the query ranges entered by a specialist.

References

1. Wei, W.: Time series analysis: univariate and multivariate methods. Pearson Addison Wesley (2006)
2. Torres, R.D.S., Falcão, A.X.: Content-based image retrieval: Theory and applications. Revista de Informática Teórica e Aplicada 13, 161–185 (2006)
3. Zhong, S., Gang, W.: Study on algorithm of dependent pattern discovery of multiple time series data stream. In: 2011 International Conference on Computer Science and Service System (CSSS), pp. 767–769 (2011)
4. Faloutsos, C., Ranganathan, M., Manolopoulos, Y.: Fast subsequence matching in time-series databases. In: Proceedings of the 1994 ACM SIGMOD International Conference on Management of Data, SIGMOD 1994, pp. 419–429. ACM, New York (1994)
5. Keogh, E.: A fast and robust method for pattern matching in time series databases. In: Proceedings of WUSS 1997 (1997)
6. Agrawal, R., Faloutsos, C., Swami, A.N.: Efficient similarity search in sequence databases. In: Lomet, D.B. (ed.) FODO 1993. LNCS, vol. 730, pp. 69–84. Springer, Heidelberg (1993)
7. Kent, A., Berry, M.M., Luehrs, F.U., Perry, J.W.: Machine literature searching viii, operational criteria for designing information retrieval systems. American Documentation 6(2), 93–101 (1955)
8. Agrodatamine: Development of Algorithms and Methods of Data Mining to Support Researches on Climate Changes Regarding Agrometeorology (2013), `http://www.gbdi.icmc.usp.br/projects/agrodatamine/index.html`
9. UCI Machine Learning Repository: Diabetes Data Set (2013), `http://archive.ics.uci.edu/ml/datasets/Diabetes`