DYNAMIC TIME WARPING FOR EFFICIENT RANGE QUERY

Chuyu Li, Long Jin, Sungbo Seo, Keun Ho Ryu

Dept. of Computer Science, Chungbuk National University
12, Gaesin-dong, Heungdeok-gu, Chungbuk 361-763, Korea
{lichuyu, kimlyong, sbseo, khryu}@dblab.chungbuk.ac.kr

ABSTRACT:

Time series are comprehensively appeared and developed in many applications, ranging from science and technology to business and entertainment. Similarity search under time warping has attracted much interest between the time series in the large sequence databases. DTW (Dynamic Time Warping) is a robust distance measure and is superior to Euclidean distance for time series, allowing similarity matching although one of the sequences can elastic shift along the time axis. Nevertheless, it is more unfortunate that DTW has a quadratic time. Simultaneously the false dismissals are come forth since DTW distance does not satisfy the triangular inequality. In this paper, we propose an efficient range query algorithm based on a new similarity search method under time warping. When our range query applies for this method, it can remove the significant non-qualify time series as early as possible before computing the accuracy DTW distance. Hence, it speeds up the calculation time and reduces the number of scanning the time series. Guaranteeing no false dismissals, the lower bounding function is advised that consistently underestimate the DTW distance and satisfy the triangular inequality. Through the experimental result, our range query algorithm outperforms the existing others.

KEY WORDS: Dynamic time warping, Time series, Range query, Similarity search

1. INTRODUCTION

Time series is a ubiquitous form of data that appeared and developed in virtually applications, such as science, technology, business and entertainment. Similarity search in time series database has attracted much interest. It is of growing importance in many applications such as information retrieval, bioinformatics, data mining and data warehousing [3,6].

Similarity search always focused on two interesting problems. The one is whole sequence matching [1,5,7,10, 11]. Another is subsequence matching [4]. Similarity measure popularly used is the Euclidean distance [1,4,9] between the subject sequences. Although it can be computed relatively fast, it is not an intuitively effective distance measure due to its sensitivity to some outliers such as amplitude scaling, offset translation and distortions in the time axis. Dynamic time warping (DTW) [2,7] is a much robust distance measure for time series, allowing similarity matching although one of the sequences can elastic shift alone the time axis. However, it is unfortunate that DTW has a quadratic time. It fails to satisfy the triangular inequality [10] so that false dismissals [1,4] come forth. To overcome these limitations, previous researches proposed some computationally cheap lower bounding functions [5,7,8,10,11] instead of the calculation of the DTW distance that can guarantee no false dismissals.

In this paper, we propose an efficient ε-range query algorithm to operate the similarity search based on [8] in that one k-nearest neighbor algorithm only was introduced. It can speed up the calculation time of the DTW distance and reduce the number of scanning the time series. Through experimental result, our range query algorithm outperforms other existing approaches.

The rest of the paper is organized as follows. In Section 2, we will consider the background and related work. Section 3 will describe the efficient ε-range query algorithm proposed by us. In section 4 we will discuss the experimental evaluation. Finally, in section 5 conclusions are given.

2. BACKGROUND AND RELATED WORK

We first summarize in table 1 a list of symbols used in the rest of paper.

2.1 Dynamic Time Warping (DTW)

The standard definition of Dynamic Time Warping (DTW) distance is as follows:

Definition 1: the DTW distance between two time series S, Q is

$$D_{DTW}(S,Q) = D_{base}(First(S), First(Q)) + min \begin{cases} D_{DTW}(S, Rest(Q) \\ D_{DTW}(Rest(S), Q) \\ D_{DTW}(Rest(S), Rest(Q)) \end{cases}$$

D_{base} can be any of the distance function defined in [10]. From the $N \times M$ matrix we can get a warping path, W, from cell $(1,1)$ to (N,M) corresponds to a particular alignment, element by element, between S and Q:

$$W = w_1, \ldots, w_k, \ldots, w_K, \quad \max(n,m) \le K < n+m-1.$$

$$w_k = (i,j), \quad w_{k+1}(i`, j`)$$

This warp path must satisfy two constraints that are continuity ($i`-i \le 1$ and $j`-j \le 1$) and monotonic ($i`-i \ge 0$ and $j`-j \ge 0$).

We can use the cumulative distance $\alpha(i,j)$ as the matrix distance $d(i,j)$ in the current cell and the minimum of the cumulative distance of the adjacent elements:

$$\alpha(i,j) = d(i,j) + \min\{\alpha(i,j-1), \alpha(i-1,j-1), \alpha(i-1,j)\}$$

Table 1. List of symbols.

Symbol	Definition
S	a data sequence
Q	a query sequence
s_i	the $i-th$ element of the S
q_i	the $i-th$ element of the Q
ε	the tolerance in range query
S^{seg}	the approximate segment sequence of S
s_i^{seg}	the $i-th$ segment in the S^{seg}
s_i^{seg}.max	the maximum value of the $i-th$ segment
s_i^{seg}.min	the minimum value of the $i-th$ segment
s_i^{TI}	the time interval of the S

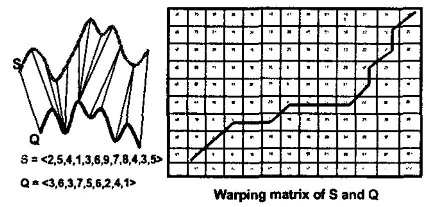

Figure 1. Illustrate the DTW. Left one is alignment of two time series according to right warping matrix.

We can find the optimal warp path according to the minimum warp cost. This process is illustrated in Figure 1.

2.2 Related Work

The database community has been researching problems in similarity query for time series databases for many years. Agrawal dt al. [1] utilized the Discrete Fourier Transform (DFT) to transform data from the time domain into the frequency domain and used a R*-tree to index the first few DFT coefficients. Yi et al. [10] first investigated the DTW in large database that used FastMap technique to approximate index the time series under dynamic time warping distance but it is not able to guarantee no false dismissal. Kim et al. [7] introduced the range query based on proposed lower bounding function that is employed on four features extracted from each time series that are first, last, greatest and smallest elements.

Keogh [5] proposed another lower bounding function and exact indexing of DTW, which was later optimized by Zhu and Shasha [11]. The proposed lower bounding function guaranteed no false dismissals. Sakurai et al. [8] proposed a new method FTW (Fast search method for dynamic Time Warping), which efficiently pruned a significant number of the search candidates to reduce the search costs. They applied a k-nearest neighbor search algorithm during the similarity query processing. Utilizing this method as a start point we propose an efficient ε-range query algorithm under the time warping distance.

3. PROPOSED METHOD

3.1 Coarsening Computation

3.1.1 Approximate Segment Sequence

Considering the time complexity of DTW $O(NM)$ where lengths of two time series are N and M respectively, it is an optimal method that we use the approximate segment sequence of the time series to compute the DTW distance between two time series especially for long sequences. First we divide one time series into several segments by the time interval (TI) and this divided time series is called the approximate segment sequence S^{seg} that is described as $S^{seg} = \{s_1^{seg},...,s_i^{seg}\}$. Every separated segment s_i^{seg} is denoted by its maximum value s_i^{seg}.max, minimum value s_i^{seg}.min and time interval s_i^{TI} as in Figure 2.

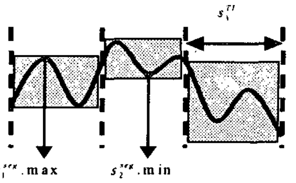

Figure 2. Approximate segment.

In terms to Figure 1 we can construct approximate segment sequences S^{seg} and Q^{seg}. We assume the time interval is 3 (that is TI =3). Instead of compute the exact DTW distance between the time series S and Q using the D_{DTW}, we can use a new lower bounding function D_{TI-LB} to calculate the approximate DTW distance of S^{seg} and Q^{seg}. Moreover the time complexity is reduced to $O(\frac{NM}{TI^2})$.

3.1.2 A New Lower Bounding Function

Definition 2: The approximate DTW distance between two approximate segment sequences S^{seg}, Q^{seg} is:

$$D_{TI-LB}(S^{seg},Q^{seg}) = D_{base}^{seg}(First(S^{seg}),First(Q^{seg}))$$

$$+min\begin{cases} D_{TI-LB}(S^{seg},Rest(Q^{seg})) \\ D_{TI-LB}(Rest(S^{seg}),Q^{seg}) \\ D_{TI-LB}(Rest(S^{seg}),Rest(Q^{seg})) \end{cases}$$

The cumulative distance $\beta(i,j)$ as the matrix distance $d^{seg}(i,j)$ in the current cell and the minimum of the cumulative distance of the adjacent elements:

$$\beta(i,j) = d^{seg}(i,j) + \min\{\beta(i,j-1),\beta(i-1,j),\ \beta(i-1,j-1)\}$$

$$d^{seg}(i,j) = \min(s_i^{TI},q_j^{TI}) \times D^{seg}(s_i^{seg},q_j^{seg}),$$

where $D^{seg}(s_i^{seg},q_j^{seg})$ denotes the distance between s_i^{seg} and q_j^{seg}. As two approximate segment sequences, the distance between every segment can be obtained by following formula:

$$D^{seg}(s_i^{seg},q_j^{seg}) = \begin{cases} (s_i^{seg}.\text{min}-q_j^{seg}.\text{max})^2 & s_i^{seg}.\text{min} > q_j^{seg}.\text{max} \\ (q_j^{seg}.\text{min}-s_i^{seg}.\text{max})^2 & q_j^{seg}.\text{min} > s_i^{seg}.\text{max} \\ 0 & (otherwise) \end{cases}$$

Theorem 1: For any two time series S and Q, S^{seg} and Q^{seg} are their approximate segment sequences, the following inequality always holds.

$$D_{TI-LB}(S^{seg},Q^{seg}) \leq D_{DTW}(S,Q).$$

Due to the limitation of the space we omit this proof here.

So we can easily derive corollary from theorem that is:

$$D_{DTW}(S,Q) \leq \varepsilon \Rightarrow D_{TI-LB}(S^{seg},Q^{seg}) \leq \varepsilon.$$

This corollary implies that similarity search that uses D_{TI-LB} rather than D_{DTW} in order to discard dissimilar sequences does not incur false dismissals.

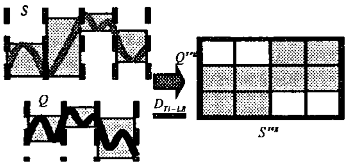

Figure 3. Calculation of the DTW distance between the approximate segment sequences S^{seg} and Q^{seg} .

```
Algorithm  $CC(S^{seg}, Q^{seg}, \varepsilon)$
// Initialize every matrix cell as  $cell(i, j)$
// $n$ and $m$ are the length of  $S^{seg}$  and  $Q^{seg}$
for  $i = 1$  to  $n$
    for  $j = 1$  to  $m$
        $cell(i, j) = \infty$;
// $wi$ and $wj$ the coordination of the element in the
//warping path;  $d(i, j)$  is the distance value of  $cell(i, j)$
//and it is computed by dynamic programming
$wi = 1$;  $wj = 1$;  $D_{seg.} = 0$;
while  $wi \le n$  and  $wj \le m$  do
    compute  $d(wi, wj)$;
    if  $d(wi-1, wj) = \min\{d(wi, wj), d(wi-1, wj), d(wi, wj-1)\}$
        $wi = wi - 1$;
    if  $d(wi, wj-1) = \min\{d(wi, wj), d(wi-1, wj), d(wi, wj-1)\}$
        $wj = wj - 1$;
    for  $i = wi$  to  $n$
        compute  $d(i, wj)$;
        if  $d(i, wj) \le (i/n) * \varepsilon$
            $cell(i, wj) = d(i, wj)$;
        else  break;
    for  $j = wj$  to  $m$
        compute  $d(wi, j)$;
        if  $d(wi, j) \le (j/m) * \varepsilon$
            $cell(wi, j) = d(wi, j)$;
        else  break;
//Compute the lower bounding distance  $D_{seg}$  of  $S^{seg}$
//and  $Q^{seg}$ , operate the range query.
    $D_{seg} = D_{seg} + d(wi, wj)$;
    if  $D_{seg} > \varepsilon$
        break;
    $wi = wi + 1$;  $wj = wj + 1$;
if  $D_{seg} \le \varepsilon$
        return  $D_{seg}$;
else return  $null$;
```

Figure 4. Coarsening Computation algorithm.

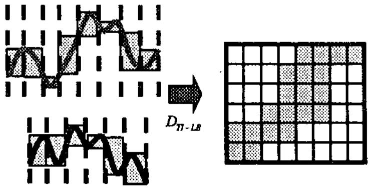

Figure 5. Farther refine the time series and calculate the approximate distance between S^{seg} and Q^{seg} .

```
Algorithm TI-RangeSearch (Q, $\varepsilon$ )
// ResultSet stored the candidate data sequence.
ResultSet = { }; compute  $A[Q]$;
for each S  $\in$  Database
    compute  $A[S]$;
    for  $i = x$  to 1
        $D_{seg} = CC(S^{seg}, Q^{seg}, \varepsilon)$;
        if  $D_{seg}$ =  $null$
            break;
        else if  $i = 1$
                put S into the ResultSet;
// Post-processing step
for each S  $\in$  ResultSet
    if  $D_{DTW}(S, Q) > \varepsilon$
        remove it from ResultSet;
return ResultSet;
```

Figure 6. TI-RangeSearch algorithm

3.1.3 Coarsening Computation

To compute the DTW distance, every cell of the matrix must be filled. Through reduce the calculation of the matrix cells we can speed up the computation of the distance. During the process of finding the warping path we can get every element in this warping path and its coordinate are wi and wj respectively. Beginning from $cell(1,1)$, every element of the optimal warping path can be got in terms of DTW definition and proposed the new lower bounding function. As soon as getting a suitable element, we can fill those matrix cells that start from the coordinate of this element and lie in the row and column. After filling them, the computed distance value of every cell is compared with the value $(i/n)*\varepsilon$ (along the wi) and $(j/m)*\varepsilon$ (along the wj) to determine whether all of the cell values are computed. In the Figure 3, due to $cell(1,2) \ge (2/3) \times \varepsilon$ we will not consider matrix $cell(1,3)$ that is illustrated using white cell. As a result we can get some white cells in one matrix that do not need to compute. During excluding the cells we also compute the approximate distance between two approximate segment sequences using the optimal warping path. In this process, we operate the range query to confirm whether two sequences are similar simultaneously. It can save a lot of computed time that we do not need to compute whole distance but partly distance. We call this calculative process and range query as coarsening computation. Our algorithm in the coarsening computation process (called CC algorithm) is described in the Figure 4.

3.2 Refinement

In this paper we regard the time interval as the division granularity. We always get the time interval that is as larger as possible. Then according to different refinement the time interval will be decreased gradually. We can get other time intervals are as following:

$$N > TI_x > TI_{x-1} > ... > TI_1 > 1$$

We propose a data structure to store every time series in the database. This data structure is a simple array that stores the information of every approximate segment sequence divided by different time intervals. Hence the S and Q are described as following:

$$A[S] \; = \{(S_1^{seg}, TI_1), (S_2^{seg}, TI_2), \; ..., (S_x^{seg}, TI_x)\}$$

$$A[Q] \; = \{(Q_1^{seg}, TI_1), (Q_2^{seg}, TI_2), \; ..., (Q_x^{seg}, TI_x)\}$$

After coarsening computation process is finished we begin to the refinement process. Followed the Figure 3 if the data sequence satisfies the condition $D_{seg}(S^{seg}, Q^{seg}) \le \varepsilon$, we will continue to refine this sequence using smaller time interval (see Figure 5) and calculate the approximate distance between two approximate segment sequences. Then we operate the range query to verify whether $D_{seg}(S^{seg}, Q^{seg}) \le \varepsilon$ is correct. Figure 6 describes our ε -range query algorithm, TI-RangeSearch, which uses the array data structure.

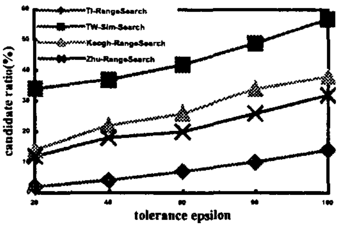

Figure 7. Candidate ratio using the Sunspot data set.

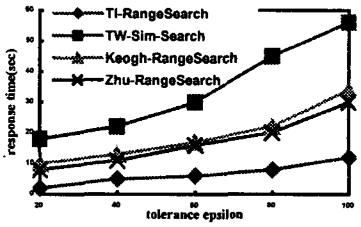

Figure 8. Response time using the Sunspot data set.

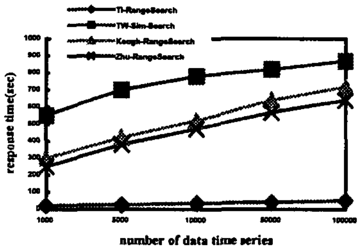

Figure 9. Response time using synthetic data set.

4. EXPERIMENTS

We applied two kinds of date set in our experiment: Sunspot data set and synthetic data set. Sunspot data set that is solar circles in the sunspot numbers every month from January 1749 to July 1990 comes from http://xweb.nrl.navy.mil/timeseries/multi.diskette. To evaluate the superior performance of our range query algorithm, we compared TI-RangeSearch that are indexed by the array with the previous ones: TW-Sim-Search [7], Keogh-RangeSearch [5] and Zhu-RangeSearch [11].

4.1 Candidate ratio and Response time

Candidate ratio is the important indicator of the filtering effect of all kinds of range query algorithms.

$$candidate\ ratio = \frac{the\ number\ of\ candidate\ time\ series}{the\ number\ of\ data\ time\ series}$$

This ratio is the smaller the better. Our first experiment compared the filtering effect of the four methods using the Sunspot data set. Figure 7 described the results of the candidate ratio of four different methods. Our TI- RangSearch outperforms other three methods.

Candidate ratio and response time are two indivisible parts for overall experimental performance. Figure 8 showed the response time of the four methods for the Sunspot data set. Our method is better than other methods. As a result, for overall performance the TI-RangeSearch is best one among all methods.

To verify the scalability of all methods, we used the synthetic data set in the following experiment due to not large enough Sunspot data set. Our experiment increased the number of the synthetic data set from 1,000 to 100,000. Under the fixed tolerance ε, the experimental result was described in the Figure 9.

5. CONCLUSIONS

In this paper, we proposed an efficient ε -range query algorithm under the time warping distance applied to [8]. Used our algorithm the non-qualify time series are pruned as early as possible before computing the exact DTW distance. The computation time and quantity of time series scanned in the large time series database are significantly reduced during the calculation of the DTW distance. For the new lower bounding function it satisfied the triangular inequality, guaranteed no false dismissal and was proved using one new method. Moreover the similarity search can be applied between the different length time series.

Through the experimental results our range query is superior to other existing approaches. In the future, we plan to further find a new method to calculate the DTW distance that is suitable with our range query algorithm.

Acknowledgement. This work was partially supported by ETRI (Telematics & USN Research Division) in Korea.

REFERENCES

[1] R. Agrawal, C. Faloutsos, and A. Swami. Efficient similarity search in sequence databases. In *Proc. of FODO*, pages 69-84, 1993.

[2] D.J.Berndt and J. Clifford. Finding patterns in time series: A dynamic programming approach. In *Advances in Knowledge Discovery and Data Mining*, pages 229-248. AAAI/MIT, 1996.

[3] M. S. Chen, J. Han, and P. S. Yu. Data mining: An overview from database perspective. *IEEE TKDE*, pages 866-883, 1996

[4] C. Faloutsos, M. Ranganathan, and Y. Manolopoulos. Fast subsequence matching in time-series databases. In *Proc. ACM SIGMOD*, pages 419-429, 1994.

[5] E. J. Keogh. Exact indexing of dynamic time warping. In *Proc. of VLDB*, pages 406-417, 2002.

[6] J. S. Kim, H. G. Lee, S. B. Seo and K. H. Ryu. CTAR: classification based on temporal class-association rules for intrusion detection. In *Proc. WISA*, page 84-96, 2004.

[7] S. W. Kim, S. Park, and W. W. Chu. An index-based approach for similarity search supporting time warping in large sequence databases. *ICDE*, pages 607-614, 2001.

[8] Y. Sakurai, M. Yoshikawa and C. Faloutsos. FTW: Fast similarity search under the time warping distance. In *Proc. of PODS*, 2005.

[9] S. B. Seo, L. Jin, J. W. Lee, K. H. Ryu. Similarity pattern discovery using calendar concept hierarchy in time series data. In *Proc. APWeb*, pages 565-571, 2004.

[10] B. K. Yi, H. V. Jagadish, and C. Faloutsos. Efficient retrieval of similar time sequences under time warping. In *IEEE ICDE*, pages 201-208, 1998.

[11] Y. Zhu and D. Shasha. Warping indexes with envelope transforms for query by humming. In *Proc. of ACM SIGMOD*, pages 181-192, 2003.