DTW

Searching and Mining Trillions of Time Series Subsequences under Dynamic Time Warping

- 整理：胡盼盼 2019/7/7
- 参考资料

http://www.cs.ucr.edu/~eamonn/UCRsuite.html
http://baike.baidu.com/view/1647336.htm
https://www.cnblogs.com/xingshansi/p/6924911.html
http://www.cnblogs.com/luxiaoxun/archive/2013/05/09/3069036.html
https://pdfs.semanticscholar.org/1116/9dec03bfd9aa5e798b9874073425a53053d2.pdf
https://blog.csdn.net/Orange_Spotty_Cat/article/details/80312154

- 组成部分
  - DTW相关基础概念
  - 论文部分

- 时间序列（Time Series）

简称时序，也就是按相等的时间采样的数据点构成的序列
轨迹(trajectory)

- 例子

心电图、脑电图、股票波动图等

- 任务

相似性搜索，给定一个时序查询Q，然后从一个时序数据库中返回与Q最相似的时序。

- ED算法

Definition 1: A *Time Series T* is an ordered list: $T=t_1,t_2,\ldots,t_m$.

While the source data is one long time series, we ultimately wish to compare it to shorter regions called *subsequences*:

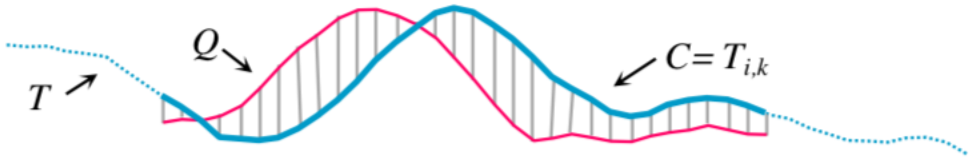
Definition 2: A *subsequence* $T_{i,k}$ of a time series T is a shorter time series of length k which starts from position i. Formally, $T_{i,k} = t_i,t_{i+1},..,t_{i+k-1}$, $1 \le i \le m-k+1$.

Where there is no ambiguity, we may refer to subsequence $T_{i,k}$ as C, as in a Candidate match to a query Q . We denote $|Q|$ as n.

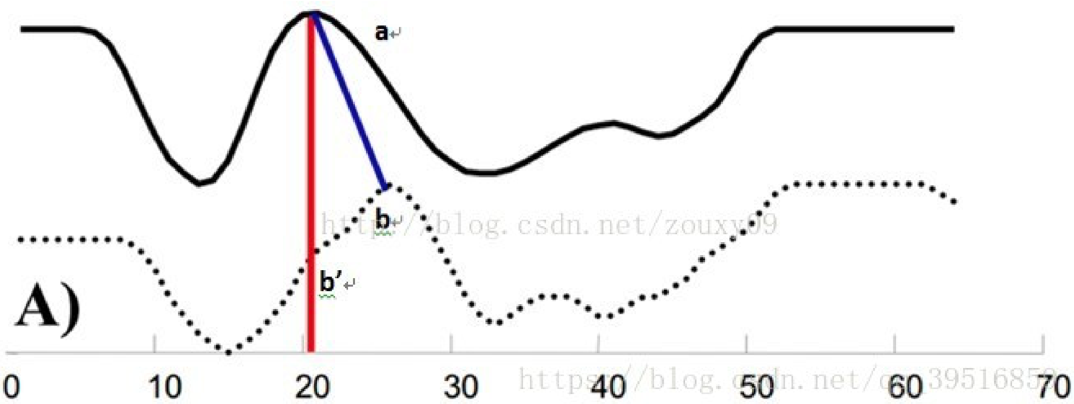
Definition 3: The Euclidean distance (ED) between Q and C, where $|Q| = |C|$, is defined as:

$$ED(Q,C) = \sqrt{\sum_{i=1}^{n}(q_i - c_i)^2}$$

We illustrate these definitions in Figure 2.

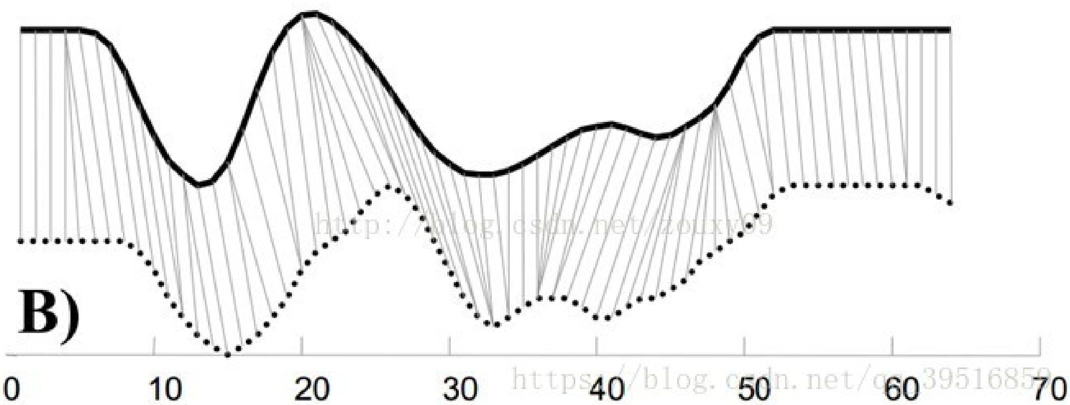


- 时间维度上的不一致性



- Dynamic Time Warping（DTW）动态时间规整

因此我们在比较序列相似性之间，先对要比较的点作一些处理，将两个序列上的点在时间轴上进行一一对齐。



- DTW例子

假设我们有两个时间序列Q和C，他们的长度分别是n和m

Q = q1, q2,…,qi,…, qn ;

C = c1, c2,…, cj,…, cm ;

d(qi,cj)=(qi−cj)^2

- 图示

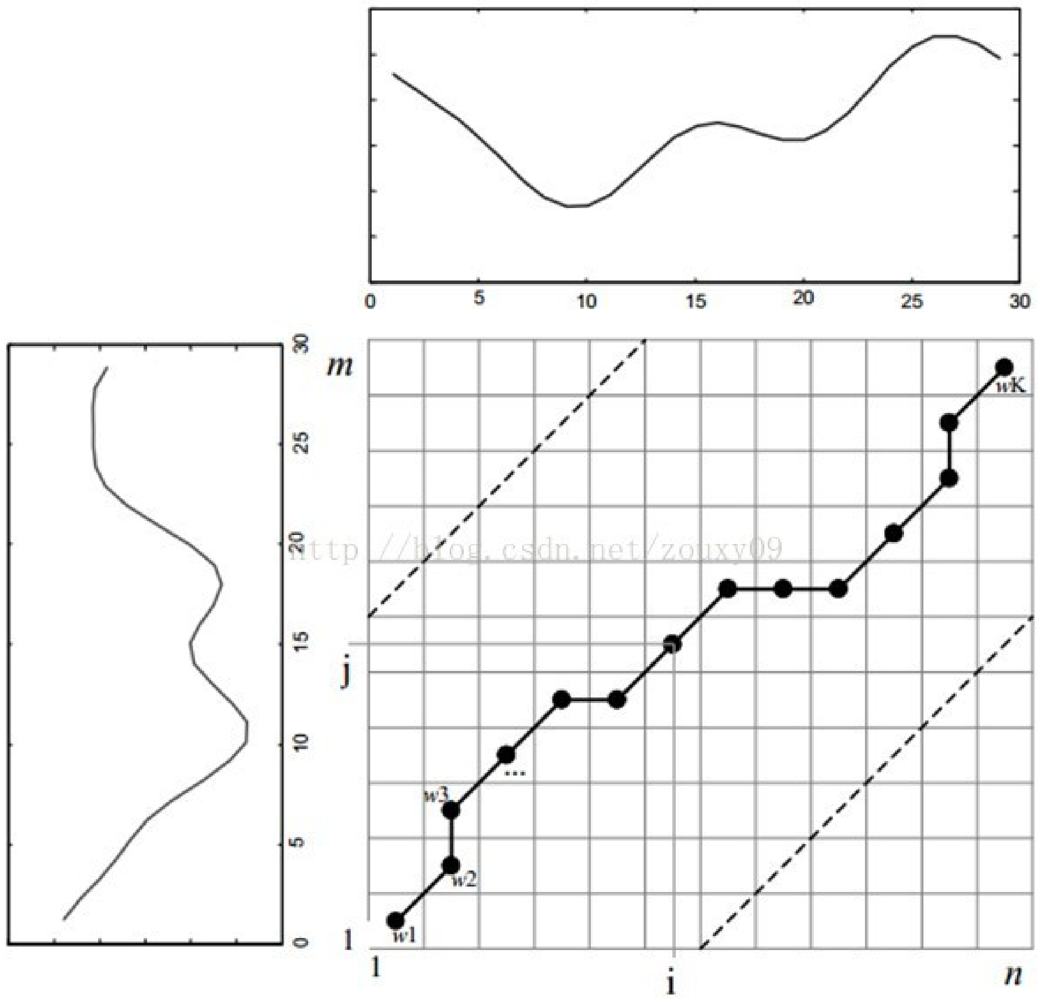
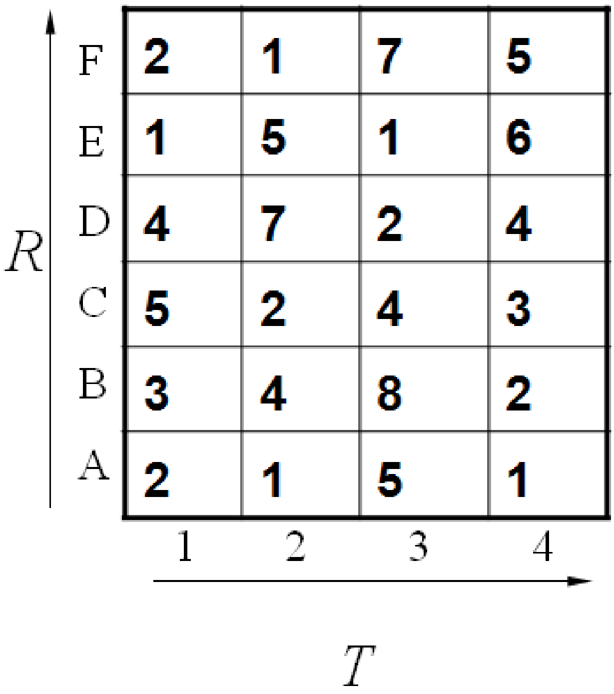


Figure 3: An example warping path.

- 动规实现

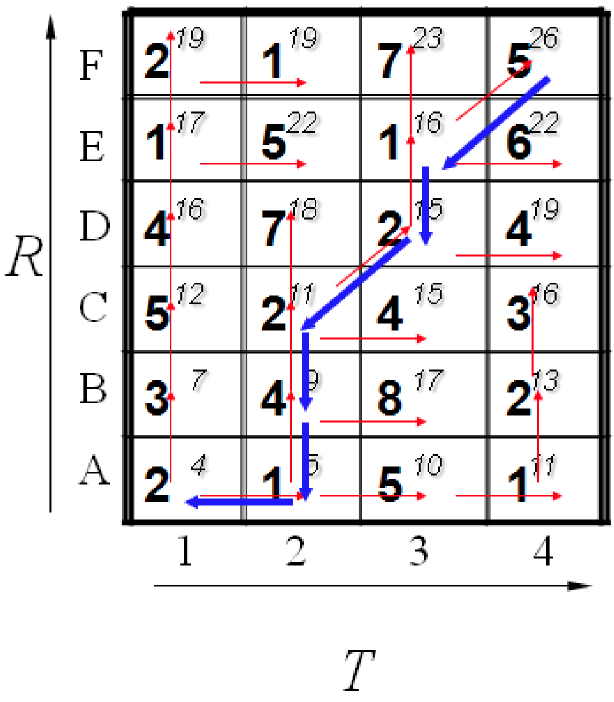


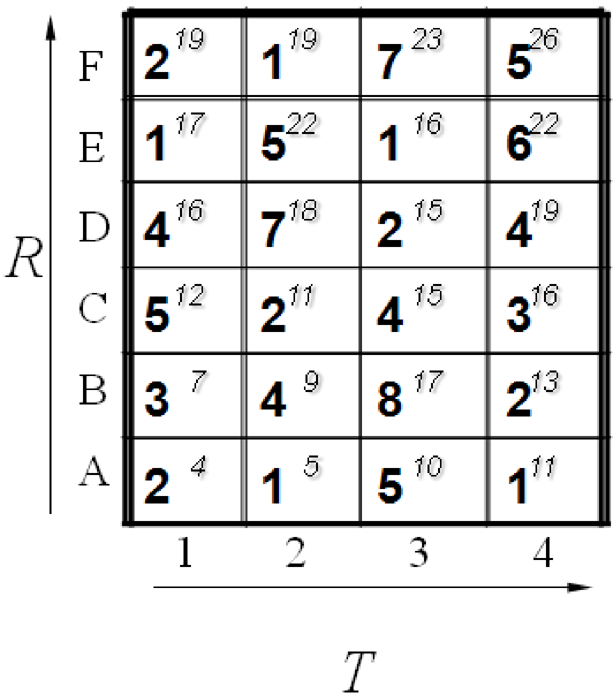
- 确切点

1）边界条件：w1=(1, 1)和wK=(m, n)是确定的。
2）连续性：路径需要是连续的。
3）单调性：也就是路径是要往前走的。

动态规划关系表达式：

$$g(ck)=g(i_k, j_k)=g(i, j)= \min \begin{cases} g(i-1, \ j)+ \ d(i, j) \\ g(i-1, j-1)+2d(i, j) \\ g(i, \ j-1)+ \ d(i, j) \end{cases}$$





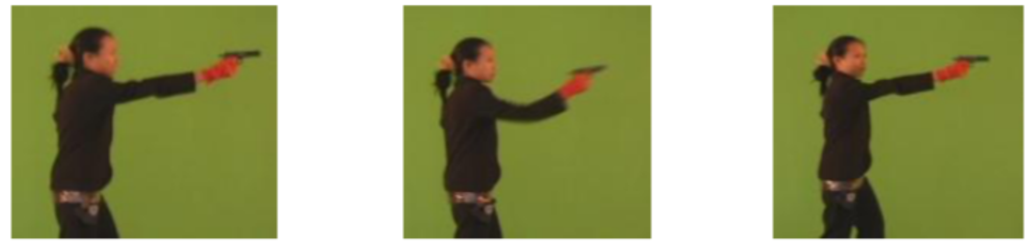
- 论文要解决的问题

应用DTW的算法复杂度太高，怎么改进?

- 概况
  - 万亿级别的数据
  - 速度极快

Likewise, a recent paper that introduced a novel inner product based DTW lower bound greatly speeds up exact subsequence search for a wordspotting task in speech. The authors state: “the new DTW-KNN method takes approximately 2 minutes” [41]; however, we can reproduce their results in less than a second. An influential paper on gesture recognition on multi-touch screens laments that “DTW took 128.26 minutes to run the 14,400 tests for a given subject’s 160 gestures” [38]. However, we can reproduce these results in under 3 seconds.

- 标准化（通过实验证明标准化之后的数据效果更好，举了一个检测枪支的例子）



- DTW是最好的方式（通过作者改进的DTW运算速度非常高）
- 添加索引任意查询长度不现实？

- 研究历史中的技巧
  - Using the Squared Distance

$$ED(Q,C) = \sqrt{\sum_{i=1}^{n}(q_i - c_i)^2}$$

  - Lower Bounding
    - Lower bounding distance
    - Best-so-far distance
    - 目标：打到一个技术差不多的人；DTW：打11局 LB：打三局
    - 计算量与筛查细致程度

    - LB_kim

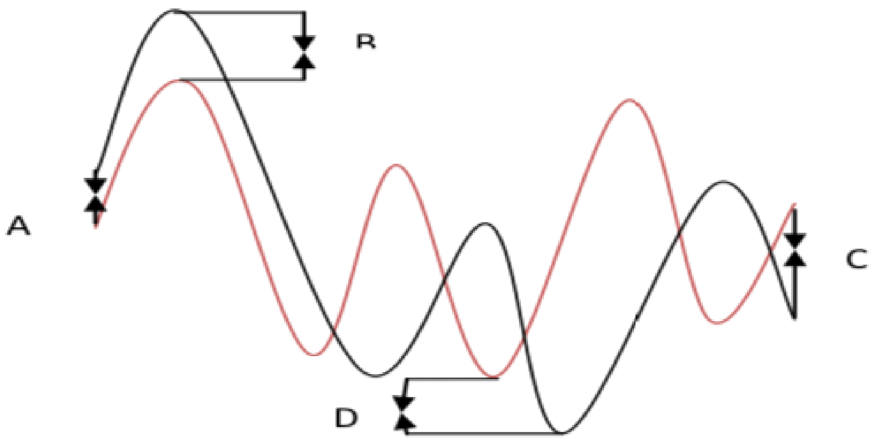
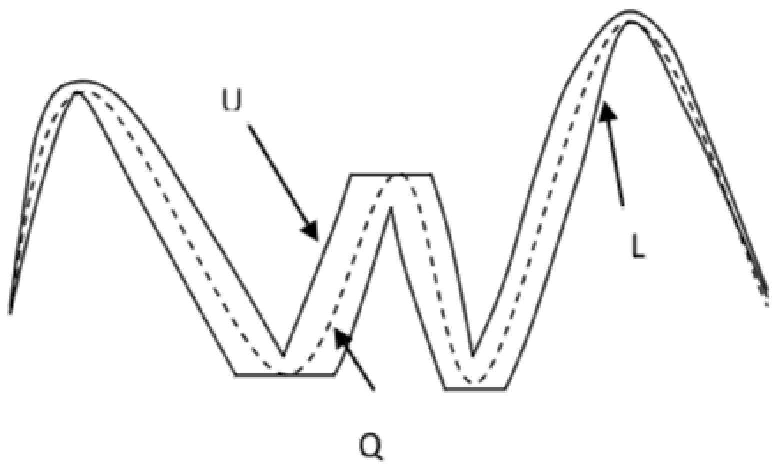


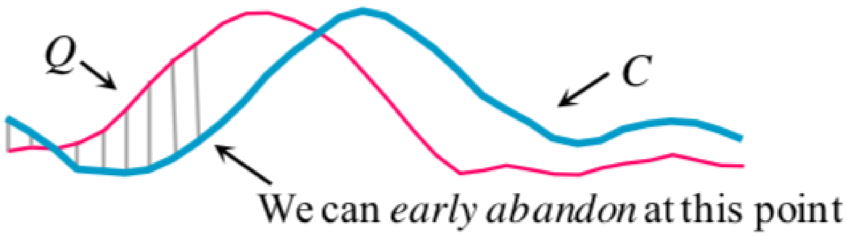
Fig.7. Lower bound by Kim et al

    - LB_Keogh

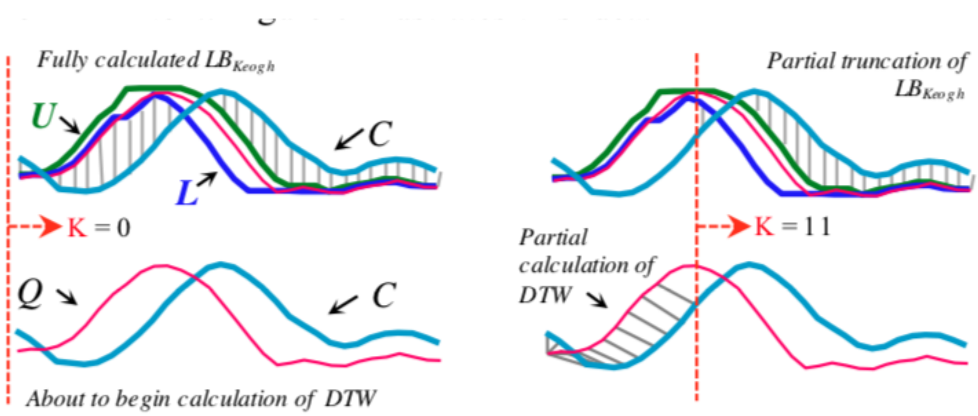


- Early Abandoning of ED and LB_Keogh

EA



- Early Abandoning of DTW



- Exploiting Multicores

* 本论文中所应用的四大技巧
  - Early Abandoning Z-Normalization

计算平均值与方差：

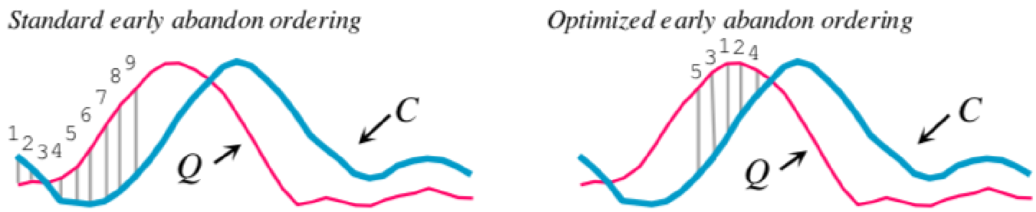
$$\mu = \frac{1}{m}\sum x_i \qquad \sigma^2 = \frac{1}{m}\sum x_i^2 - \mu^2$$

对每个点的计算：

$$\frac{x-\mu}{\delta}$$

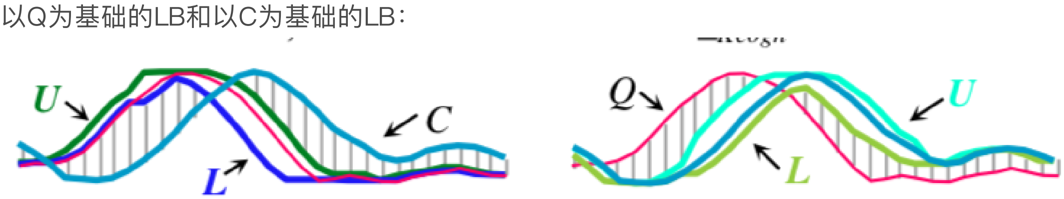
- Reordering Early Abandoning

We conjecture that the universal optimal ordering is to sort the indices based on the absolute values of the Z–normalized Q. The intuition behind this idea is that the value at Qi will be compared to many Ci’s during a search. However, for subsequence search, with Z–normalized candidates, the distribution of many Ci’s will be Gaussian, with a mean of zero. Thus, the sections of the query that are farthest from the mean, zero, will on average have the largest contributions to the distance measure.



  - 推挡、抽、拉、搓、发上旋球、发下旋球…
- Reversing the Query/Data Role in LB_Keogh
- Cascading Lower Bounds

以Q为基础的LB和以C为基础的LB：



多种LB方法的综合使用：