## 《2015 高效的时间序列对齐》小结

这个文档是对《2015 Efficient subsequence alignment of time series》的快速总结，还有我阅读时存在的问题、一些有关的想法、一些疑问等。

我拿到这篇资料有一段时间了，到现在还没有把它读完……遇到了一些问题（在下面会提到），不过我还是想先做一个小结来梳理清楚思路。

首先是先分析整个文章的结构，文章共147页，大纲如下：

1. ： 介绍（introduction）
2. ： 时间序列挖掘
3. ： UCR-Suite的并行化
4. ： 线性时间复杂度的弹性时序对齐（Towards Elastic Subsequence Alignment with Linear Cost）
5. ： Lie-Group-Valued 时间序列
6. ： 总结和未来工作

一开始我扣得很细，但是发现读不下去，总被一些难点卡住。所以后来我改变了策略，先大致搞明白各个部分讲了什么问题，梳理整体概念，遇到问题后再去找相应的章节。

以下是我对各章的理解：

### 第一章：介绍

先定义了什么是时间序列。

介绍了时间序列挖掘问题可以分为三个阶段：

1. 找到一个合适的表示
2. 确定一个相似性的概念
3. 高效且鲁棒算法的实现

介绍了两个“Toy Dataset”，和它们的性质：

1. Cylinder-Bell-Funnel (The CBF dataset)：一个人造生成的数据集
2. Gun-Point : 一个现实世界的数据

最后介绍了论文的组织。

### ：时间序列挖掘

这章中我的主要关注点是：欧氏距离（ED）、DTW、CDTW、正规化（z-normalization），除此之外还讲了很多东西（如下），但是没有特别关注。

这章首先讲了**Metrics**（度量）和 **Dissimilarities**（不相似度），我之前一直认为它们指的是同一件事情（两个事物的相似度测量），但在严格定义下它们还是有较大差别的，**比如DTW不是一种Metric（度量）(不满足一致性、三角不等式)。**

本章首先定义度量空间(metric space)[一致性、对称性、三角不等式]

还有一些命题（proposition）[非负性、倒三角不等式]

在范数族小章中定义了 normed vector space [确定性、同质性、三角不等式]，由此引出欧式距离(ED)。比较了L1范数和L2范数，并举例利用倒三角不等式来加速1-NN搜索。

接下在Pearson’s Correlation Coefficient(PCC)小章中定义了z-normalization，并应用到ED上，并证明了 z-normalization ED 和 Pearson distance平方根成正比，**且都不是一种Metric（度量）(不满足一致性)**

然后介绍了DTW算法 [单调性、连续性、边界性][非负性、对称性，不满足一致性，不满足三角不等式，所以不是一种度量metric]

然后介绍了CDTW算法，一种受限制的DTW，这个算法在后面章节会用到。

然后介绍了一些其他的相似性测量（SimilarityMeasures）、Unitary Bases、PCA、Discrete Wavelet Representations、离散傅里叶变换等，没有仔细阅读这些。

### ：UCR-SUITE的并行化

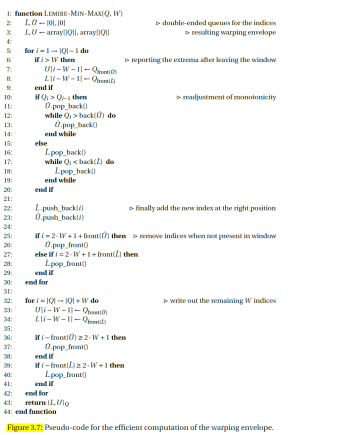
本章的目标就是**如何高效计算 ED 和 DTW**。

因为 ED 和 DTW 在各行各业广泛是用，需要高效实现，通过三个策略[并行化、开发冗余、下界]

整章分为三个副章节：

1. 介绍 The UCR-Suite (加州大学河滨分校想出来的一种高效计算的方法)

文章中用了大量的篇幅来介绍遇到的各种问题和各种解决方法，最终得出的一个算法，可以用下图来描述：



里面有许多细节问题，但核心思想就是利用下界来剪枝（提前淘汰）

1. 在 CUDA-enabled 加速器上的并行化计算

先介绍了一些硬件知识（在显卡上并行化计算的原理，我跳过了）

然后通过并行化得到了比UCR-ED 和 UCR-DTW 更快的实现

1. 性能的评估

### ：线性时间复杂度的弹性时序对齐（Towards Elastic Subsequence Alignment with Linear Cost）

本章先讨论两个传统的算法 [ FastDTW(FDTW) 和 GreedyDTW(GDTW) ]

然后介绍了一种时间序列的子序列对齐的相似度测量概念[ Gliding Elastic Match (GEM) ]

### ：LIE-GROUP-VALUED TIME SERIES

关于李群的内容，我还没有开始接触

### ：CONCLUSION AND FUTURE RESEARCH

### 总结本论文的内容： 一方面，希望得到高质量的时序对齐；另一方面希望计算代价和内存代价要小。

以上是文章的大纲结构，然后是我读文章时遇到的一些问题：

首先我意识到读这篇文章的进度确实太慢了，中间还断了一段时间，我反思后觉得存在以下问题：

1. 这篇文章真的好长，而且知识点很多很杂，读起来比较吃力。
2. 遇到了很多我不掌握的概念被卡住，比如CUDA这样的GPU并行知识和一些数学变换知识。
3. 读长篇的英语文献还是比较吃力的。有不少概念难以理解，比如“Lower Bound Cascades”中的Cascades应该怎么理解，一直不太确定。
4. 感觉缺少了些方向感，不太清楚哪些知识对我来说是重要的，哪些是可以跳过的，希望师兄能够指点~

### 接下来的打算：

接下来空闲时间会比之前多一些，

关于这篇文章，我接下打算先仔细抠一下并行化中提到的算法，

然后有重点地阅读FDTW、GDTW、GEM 这部分内容。

接下来要不要阅读“LIE-GROUP”这一章，就听师兄的建议（听到“LIE-GROUP”有点害怕，完全陌生的概念……）。

关于[快速傅里叶变换、离散傅里叶变换]的知识我还欠缺，打算找机会补上。

另外，师兄规划带我12月底投稿一篇数据挖掘期刊类期刊（DTW），我接下打算全力配合这件事情。现在不太清楚具体要做哪些工作，希望师兄指点一下，我会尽快有重点地去准备。

另外我还有一个疑问就是，之前师兄说关于这篇论文要做三个ppt，我想知道是用于讨论班分享的那种ppt吗？ 因为本科生分享是很长时间才轮一个轮回的，如果是用于讨论班，我就不急于一下子把三个都做出来，并把内容控制在一次分享的量级。