

陈纯院士：数据在时间概念下完成驱动，时序大数据正成为研究重点

AI 报道 2019-09-26

数据，看不见、摸不着，但在日常生活中无处不在。随着信息采集、处理技术的深入推进，数据已上升成为国家基础性战略资源，有人将大数据比喻为 21 世纪的金矿石。大数据与普通数据相比，最大的特点是带有时间戳，所以也被称作**时序大数据**。



未来，基于时序大数据的实时智能技术具有广阔应用前景。那么，时序大数据会应用在哪些场景？又将怎么样发挥出数据的价值？在 2019 上

海静安国际大数据论坛上，**中国工程院院士、浙江大学信息学部主任陈纯就“时序大数据实时智能：技术及应用”进行了分享。**



中国工程院院士、浙江大学信息学部主任 陈纯

时序大数据实时智能：技术及应用

身处数据驱动的大数据时代，如何变成一个**大数据的强国**，**技术**在其中是非常重要的。

首先，为什么有那么多数据？其实，在互联网以前，我们只有人类社会和物理世界，然后才有的数据产品以及信息空间，我们称之为 CPH。在信息空间，包括人工智能、AR、VR 都非常重要，产生了很多的数据，但并不仅仅是取量变多才是大，更重要的是**把历史的数据都留下来了**。

每一个数据都是带有时间的，以前的特征向量、特征空间把时间去掉了，只能挤在一起成为扁平的数据。移动互联网和物联网使得每一个数据都有时间戳，带有时间的数据可以做很多的处理，尤其是 **5G 的到来**，对热数据的处理非常重要。

另外一个就是数据形成**关系图谱**、**关联图谱**，就像社交，以前的关联图谱不带有时间，而基于时序大数据，以前的数据得以留存，现在的数据也时刻流进来。**时序大数据**是以前的历史数据，加上实时的数据，这好像水库的水，正在流向水库里的水是流数据，留在水库里的水都带有时间。实时的数据可以称之为热数据，慢慢的变成温数据，然后冷数据。**时序大数据非常重要，而且是最近几年研究的重点**。

数据随着时间的推移会越来越不值钱，所以要把它处理起来。直到 **2012 年**，**实时大数据才被重视起来**。但现在的互联网公司大多仅仅是当时的数据，没有把历

史数据加以考虑，还停留在以前常用的大数据分析方式。实时的数据仅仅是流，没有把以前的数据结合起来考虑，这是很浪费的，并且恰恰很多的应用需要把历史数据考虑进来。**2014 年**的时候，才有**时序数据的相关概念**出来。很多算法有非常多讲究，要做到大数据量，而且要尽量低的延时，这需要有很多的技巧。**2018 年**，才有了**分布式实时时序图数据**。

时序大数据实时分析的关键技术在哪里，为什么这么难？

关键技术一：面向复杂统计指标的增量计算。大数据的分析，一些统计指标特征计算是非常重要的，均值、方差等等。简单算法、静态取数、容器类算法、复杂算法、CEP 等分别如何实现？增量计算中如何进行退单等常见场景的逆向计算？事件乱序抵达如何确保增量计算的结果一致？这些数理统计算法中的增量计算、可逆计算、乱序计算等问题需要考虑。

关键技术二：面向时序数据处理的动态时间窗口。时间窗口需提供滚动、滑动的漂移能力，支持长周期时间窗口的动态精度控制，支持基于弹性时间窗口的实时 ADHoc 查询。

关键技术三：基于流的事件序列识别（复杂事件处理 CEP）。事件模式的增量匹配、叠加通用算法的增量统计等支持 CEP 的增量匹配及数理统计问题。

关键技术四：动态时序图谱的实时分析计算。时序图谱的极速增量建图，大规模时序图谱如何提供百万 tps 的建图能力？时序图谱的分布式处理，10 亿顶点，100 亿边(10 亿时序复杂边)的前提下，3 层以上查询如何控制在秒级？大规模时序图谱如何秒级地图搜索(最短路径、Page Rank、Louvain、LPA 等)能力？面向时序图谱的查询语言。支持动态时序图谱的时间维度 Ad Hoc 查询分析能力？比如说中国移动有 9 亿个电话号码，在 9 亿个人的点里面，每每两个通信频里最频繁的 100 个人，能不能找出来？如果实时建个图的话，就有 9 亿个点，每每两个可能有 100 个关系。按照现在的来做的话，可能要找几个小时，甚至几天时间，但用时序图可能几秒钟就出来了。

实时的数据怎么形成一个智能平台？

只要是智能的平台，一定会有所谓的**智能模型**。智能模型不仅仅指深度学习，深度学习最大的贡献之一是它能利用大数据（批式、标记）进行训练，从而获得多层次的数据特征，利用这些特征能大大提升模型对数据的分类精度。分析计算+智能模型，便构成了时序大数据实时智能技术架构，可以进行实时采集、实时加工、实时分析、实时决策。

实时分析对于很多行业都是重要的，目前流立方（图立方）时序大数据实时智能处理平台已在 400 多家单位得到成功应用。以上海的银联为例，只要一刷卡，银联会判别这张卡是不是本人行为，是不是被盗了。银联需要在刷卡的时候有数据才能进行后续的判断，这个数据怎么来的呢？比如说，用户在纽约刷卡，可能有

20 个特征的向量数据传到了上海，包括身份证号码、卡号、消费地点、消费金额等等，数据传到后台，银联马上要计算用户的统计指标。哪些数据会有用？用户平均的消费均指、在不同场合的消费方差等等，可能要计算二十几个统计指标，把用户五年来所有的销售，全部统计起来，计算好这些指标，然后把这个指标送进一千多条基于模型建立的规则。比如一个简单的规则，上次的消费是在上海，但是这次消费是在纽约，相隔时间只有 6 个小时，那用户的卡肯定出问题了。这些几千条的规则算出来的，银联要求每秒 5 万笔，实时的判断，20 毫秒以内判别，用流力方和银联专家风控的规则结合起来，就形成了一个非常好的规则。

综合来看，第一，**实时计算**非常重要，数据驱动一定是随着时间来驱动的，需要把历史数据和实时数据综合考虑进来。第二，**实时的智能系统**可以帮助专家或从业人员基于应用场景把模型做的更好。

我们认为，从明年开始，甚至从今年开始，在**时序大数据实时智能技术及应用领域**，应该有很多人在研究，这也是当下需要做的事情。