Bigdata 17

真值发现现状：假定待考察的变量的值不随时间变化（静态真值发现）或假定MV的value随时间变化的行为与MV本身独立（针对动态而言）。

存在的问题：（1）缺少对于MV的物理约束，即：例如在恐怖袭击事件中，随时间的变化，伤亡人数不可能下降。（2）数据存在噪声或者不完整的问题，并且对于可信数据源的确认也存在问题。例如报道袭击事件，一些大型媒体例如CNN报道错误，但是online媒体例如tweet等，在评论中或者发tweeter中，存在着正确信息。一方面，数据中的噪声可能招致数据的错误，使得对于MV的真值发现，数值转换出现错误；另一方面，数据的不完整对于动态真值发现问题的证据可信度提出了挑战。

问题重述方面，对于V，MV，约束进行了形式化定义，约束定义为硬性条件约束（在任意时间点遵守）和软约束（描述了MV真值在一个时间点转化更新的可能性）。对于数据源，本文采用基础数据源primary：S和互补数据源（complementary source，为了处理具有噪声的数据所提供的数据源），

问题解决方法：总体来看，本文提出了CA-DTD这一基于约束注意机制的动态真值发现技术。他有CA-HMM（经过改造的隐式马尔可夫模型）以及CSI（一种将数据源的数据进行数据融合的操作）。其中，CAHMM用于在观察来自社交媒体的数据后，从其中待观察的MV中捕捉真值。与传统的HMM模型不同的是，它通过将物理约束融入隐式马尔科夫过程的状态转换规则，也就是改造Viterbi decode过程，从而加入了对于物理约束的注意机制，实现对于上述问题一的解决。

另一方面，CSI组件用于融合互补数据源中的一些规则（claim）从而和我们改造后的CAHMM模型相融合。这种互补的数据源（traditional news media），可以提供的基础数据源所无法注意到的信息。进一步对于问题二中数据噪声吗，不完整等问题进行了补充和数据融合修复。

对于隐式马尔科夫过程，我们定义的隐藏层是无法被直接观察到而地真值；反之，从数据源中可以直接观察到的（这里的数据源是通过CSI过程计算，融合过后的）形成了ACO（crowd opinion）。该模型的输入是一系列对于变量的观察值ACO，输出是一系列的与MV相关的真值估计序列。

注意：对于约束的设定，是基于物理定律和先验知识的！

对于指定的MV，其真值的转化难度本文给出了一个量度：HA

对于物理约束，我们分为如下四类：

（1）硬约束1：对于一系列MV，其真值必须有指定的顺序，例如伤亡人数单调不减；

（2）硬约束2：在某一特定的时间点，某些状态可能不可达。例如在事故发生的半小时内，我们认定，犯罪者不可能逃离犯罪区域200公里以上。

（3）软约束1：状态转换的频率不应超出一个阈值。

（4）软约束2：状态转换应该遵循时间和空间的约束。

特别的，在指定的约束下，对于一个MV，真值转换的难度我们给出了一个HA函数，后续会用到。

对于传统Viterbi decode过程的改进：这一decode过程的作用是，在我们给定的物理约束下，寻找最可能产生我们能够观测到的，经过融合了的ACO的，真值序列。进一步的，提出了AHC score去衡量这一转化过程，这里就用到了之前提到的HA函数进行计算。AHC score用来判定这种转换是否被允许。在这一算法中，我们递归的定义了在观察到前t个观察值之后，当前值为真值的可能性。

对于数据源而言：

由数据源所作出的claim是否可信，他对于真值发现的贡献量度，用CS表示，不确定量度用k表示。Claim独立性量度independent score。这三者进行分别计算，然后共同组成了对于一个数据源的claim衡量其可信度评分trust score的三个组成部分。进一步，对于数据源给出了其权重的表示，由其MV以及ACO意见和我们估测的真值是否吻合的一致性函数组成。算法部分，先给出了改进后的decode算法，然后给出了整个CADTD的伪代码。