数据集

目前是有4组benchmark：weiboA数据集、weiboB数据集、fakenewsnet数据集、mediaeval数据集；

weiboA数据集包括推文，附加图片（多张），标签。没有时间戳，没有对应事件，EANN提供了single-pass聚类后的事件标签，相当于进行了一次社区发现。

weiboB数据集包括推文，附加图片（多张），账户信息（性别、关注数、粉丝数、定位、身份认证），所属类别。其所属类别标签是自带的。

Fakenewsnet数据集和weibo数据集类似，包括推文、图片和标签。没有时间戳，没有对应事件。唯一区别的是推文长度非常长，和weibo的短文本对比鲜明。

Twitter

Mediaeval数据集包括推文，附加图片，对应事件，时间戳，标签。其中事件的新闻数量分布不平衡，sandy事件在训练数据集中占了75%，测试数据集较为平衡。下面是训练集的分布图，可以看出sandy、nepal、boston推文数量较多且真假平衡。

图表, 条形图

描述已自动生成

原本计划以事件、事件、数量为轴进行统计分析，后面粗略看了下，所有事件的相关推文在时间轴上并不重叠，且每个事件推文的时间一般只持续4-5天，短的可能只有一两天。虚假新闻传播在时间跨度看具有短暂性。我对训练集中的几个代表性事件sandy、nepal、boston进行时序分析。结果如下：

图表, 折线图

描述已自动生成

图 1 sandy

图表, 折线图

描述已自动生成

图 2 nepal

图表, 折线图

描述已自动生成

图 3 boston

观察这几个事件的演化规律，首先几个事件之间在时间跨度上确实没有交集，这些事件都是离散在时间轴上的独立事件，前后关联不得而知。但是在真实世界的新闻流中，一般情况是同时出现多类事件，如果我要采用迁移学习的方法，原先思路是这样的，背景是建立在某一段较长的时间内发生了K个事件，这些事件的相关推文较为平均的散落在时间范围内，将时间轴进行64/73/82开，前面作为源域，后面作为目标域。那么这样我就可以将源于的知识迁移到目标域进行检测，当然这有一个前提就是源域和目标域是具有可迁移性的，也就是说两者的知识存在区别但是相互关联，这也是需要提前测试的。

Weibo

在几个微博的开源数据集中，现在看来weiboB的利用价值和可操作性是最高的，不仅数量最多，而且还有领域标签，这样就无需利用聚类方法进行域发现，而是直接使用分类标签即可。