【这篇文章讲的事件与我们所说的事件不一样，应该是新话题的检测。】
【题目】基于新闻要素的在线新事件检测
【作者】李营那，阮彤，顾春华
【单位】华东理工大学计算机科学与工程系
【期刊】计算机应用与软件
【时间】2013 年 12 月

【笔记】
在线新事件检测的主要任务是从以时间顺序到来的新闻报道中识别出未知事件。
【方法描述】
构建基于新闻要素的报道和事件表示模型，该模型包括新闻报道地点、人物和内容等要素，使用多维要素的优越性在于可以区别相似事件；
计算各要素对应特征的相似度提供对应的相似度算法；
使用基于地理本体树的地名相似度算法计算地点相似度，使用基于维基百科的语义相似度计算方法计算报道内容之间的相似度；
为了衡量各要素的重要性，使用 SVM 模型训练得出各要素的权值；
以 single-pass 聚类算法为基础，在算法过程中不断修改事件的表示向量以防止事件中心的漂移；
使用滑动的事件窗口以减少因处理大量不活跃事件引起的时间消耗。
【关键词和话题 Topic】
以关键词为主体的信息检索技术是用户从海量信息中获取需求信息的主要途径：首先基于关键词对海量信息进行组织，用户在查找相关信息时，通过关键词进行匹配获得相关信息。然而在许多情况下，用户很难使用关键词准确地表达自己的真实意图，因此基于关键词的信息检索技术很难满足人们的需求。为此，以话题为主线对信息进行组织、然后以话题的方式把相关信息展现给用户，成为信息获取的另一种重要方式。话题是指一个种子事件或活动以及与之直接相关的事件或活动。
【新事件检测 NED】
NED 是 TDT 的一项重要子任务，NED 的目标是从时序新闻源中检测出一个新闻话题种子事件的第一篇报道。
【相关工作】
文献【2】采用 K-means 算法，将语料中的第一篇报道作为一个初始簇，其余报道按时间顺序处理，对于每个目标报道计算它和已存在簇之间的距离，当该距离大于阈值时产生一个新的簇也就是检测到一个新事件，否则把目标报道归入距离最近的簇。
文献【3】采用单路径聚类算法，当新报道达到后立即提取该报道的特征术语，建立报道内容的查询表示，然后将该查询和已存在的所有查询进行比较，如果比较结果没有超过阈值则认为检测到一个新事件。
文献【4】也采用单路径聚类算法，当新报道达到后提取报道的特征，建立报道的向量空间表示，然后计算该向量与已有的簇的质心向量的相似度，如果相似度大于阈值则将该报道归入相似度最大的簇，否则创建以该报道为种子的新簇。
文献【5】利用神经网络的思想改进聚类算法，训练了一个多层的神经网络，在新报道和已存在的簇之间出现本质不同时建立一个新簇。
文献【11】扩展了基本的增量 TFIDF 模型，新的模型包含特定的源模型，基于特定文档均值的相似度标准化技术、基于特定源对均值的相似度标准均化、基于逆事件频率的术语权重调整和文档分割。
文献【7】采用文本分类和命名实体相结合的策略进行事件检测，每篇文档用以下三个向量

表示：文档的所有词（移除停用词）组成的向量、仅包含七种 NE（人名、组织机构、地名、日期、时间、金钱、百分比）的向量、除去 NE 的词组成的向量。

【基于新闻要素的表示模型】

地点向量 Vp：报道中出现的地名；

名称向量 Vn：报道中出现的人名、机构名；

内容向量 Vg：描述具体发生的事情；

【报道模型】

利用 ctbpaser 进行分词、词性标注；

报到时间统一表示为：YYYY-MM-DD 形式；

利用地名识别算法识别地名，构建地名向量；

构建名称向量；

提取 NN 和 VV 构建内容向量；

报道模型：S=（Vsd，Vsp，Vsn，Vsg），其中 Vsi=（ws1，ws2，…，wsn），wsi 计算方法为：

$$wt_i = \partial tf_i \times \log\left(\frac{N}{n_i}\right) \quad (1)$$

其中 tf_i 是 term_i 在 S 中的词频，N 指在线检测过程中到目前为止已有报道的总个数，n_i 是到目前为止已有报道中包含特征项 term_i 的报道的个数。式(1)对传统的 TF * IDF 进行了改进，主要引进了特征项的位置信息。本研究发现报道的标题和关键词中的词更能体现报道的内容，因此，如果 term_i 出现在报道的标题或者关键词中，取经验值 1，否则取 0.6。

【事件模型的构建】

时间：归入该事件的最近一篇报道中描述事件的发生时间；

计算报道中特征词权重；

更新事件特征词列表；

报道的特征词在事件中的权值是通过式(2)计算得到的：

$$w(\delta_i, t_j, E_n) = \frac{\sum_{S_{mj} \in E} W_{tj}(\delta_i, S_{mj})}{N_E} \quad (2)$$

其中 $w(\delta_i, t_j, E_n)$ 表示特征词 δ_i 在 t_j 时刻在事件 E_n 中的权值；N_E 是事件 E 在 t_j 时刻所包含的报道的总数；$S_{mj}(1 \leqslant m \leqslant N_E)$ 为事件 E 在 t_j 时刻所包含的报道；$W_{tj}(\delta_i, S_{mj})$ 是特征词 δ_i 在 t_j 时刻在报道 S_{mj} 中的权值，可利用式(1)计算得到。

【报道和事件相似度计算】

1、基于维基百科的同义消解，归并同义词；

2、名称子向量与内容子向量余弦相似度计算；

3、基于地理树的地名子向量相似度计算；

$$S_p(p_1,p_2) = \frac{2deep(p_1 \cap p_2)}{deep(p_1) + deep(p_2)} \quad (4)$$

其中 $deep(p_i)$ 表示地名 p_i 在地理树中距离根节点的路径长度，例如图 2 中 $deep(\text{北京市}) = 1$，$deep(p_1,p_2)$ 则表示 p_1 和 p_2 的直接共同祖先距离根节点的路径的长度。报道中通常包含多个地名，报道和事件的地名子向量 $V_{SP} = \{x_1,x_2,\cdots,x_n\}$ 和 $V_{EP} = \{y_1, y_2,\cdots,y_n\}$ 的相似度利用式(5)计算得到。

$$Sim(S_P,E_P) = \sum_{i=1}^{N_S}\sum_{j=1}^{N_E} \frac{n_i}{C_S} \times \frac{n_i}{C_E} \times S_p(i,j) \quad (5)$$

其中 N_S 和 N_E 分别表示报道 S 和事件 E 的地名子向量的维度，C_S 和 C_E 分别表示报道 S 和事件 E 中的地名术语的总个数，n_i 表示地名 i 在报道或者事件中出现的频次，$S_p(i,j)$ 表示两个地名术语 i 和 j 利用地名本体树计算得到的相似度(可由式(4)得到)。

【新事件检测算法】

事件簇的滑动窗口：以当前时间为终点，仅检测报道是否属于窗口中的事件；

【实验结果】

语料集：新浪网 2011 年新闻频道（内地新闻）的 14322 个报道，并从对应的专题频道中经过处理得到 48 个事件。

经过训练，话题数量定位 30；

经过训练，事件/话题特征词定位 100；

经过训练，滑动时间窗口定为 8；