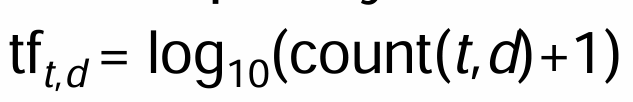
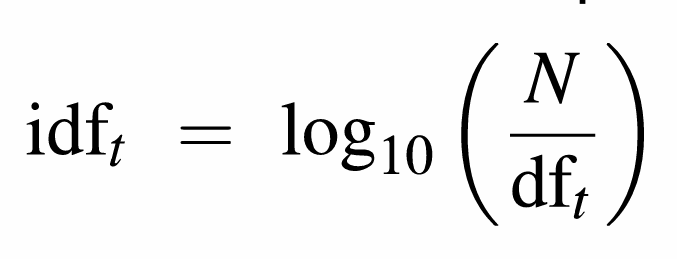
**第五篇论文《Mining the Web to Predict Future Events》读后感**

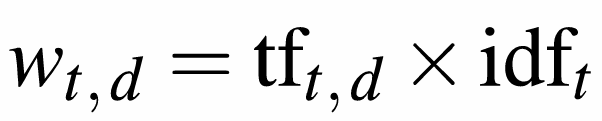
陈欣禾 2021533093

摘要：这篇文章通过分析过去事件的新闻数据来预测未来事件的发生，不仅仅预测事件是否发生，也预测发生的详细信息例如死亡人数。它的做法是首先利用故事情节挖掘和聚类，从大量新闻中提取关键信息，相比于通过使用时间上的阈值等和时间相关的方法来筛选文章，这里他使用了熵，筛选和事件相关的文章作为样本；其次，在获取词汇和事实特征的时候，使用适用于各自特征的形式去编码数据，例如在获取事实特征的时候从Wiki中获取信息给词分等级，慢慢细化；接着使用贝叶斯规则构建概率模型，假设事情T时刻的状态是由T时刻的特征组成的，并且假设特征和特征之间是独立的，通过当前时刻的特征一步步推导到下一个时刻事件的结果；最后为了去除稀疏性，通过图网络结构将事件概括为更一般的概念，就比如说在霍乱这个事件上，不再局限于某一个地点的霍乱，而是更针对于霍乱这个事件。

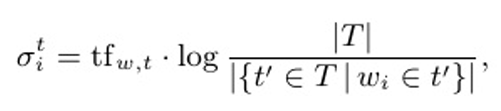
第一个想法：在故事情节挖掘和聚类对的时候，他使用了修改的Inc.LM方法，能够更准确地识别关键文章，我认为有点像TF-IDF （Term Frequency-Inverse Document Frequency）。TF-IDF通过统计TF和IDF来评估一个词对于一个文档集合的重要程度。其中TF（词频）指的是一个词在文档中出现的频率，一个词在文档中出现的次数越多，它对应的TF值就越高。IDF（逆文档频率）衡量一个词的普遍重要性，如果一个词在很多文档中都出现，它的IDF值就较低；如果一个词只在少数文档中出现，它的IDF值就较高。



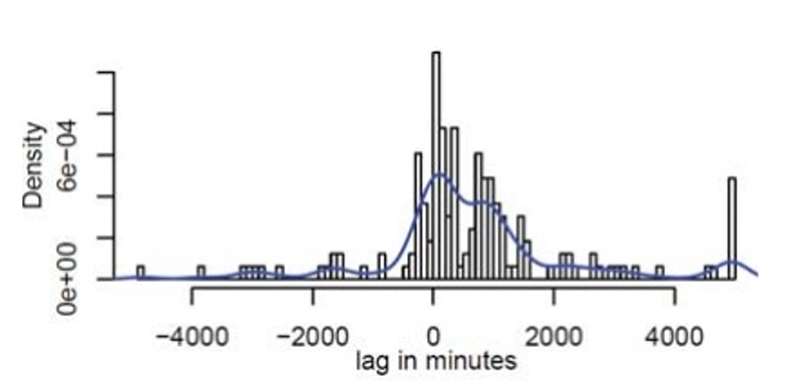




可以看到文章中的这个式子和TF-IDF没有区别——其中|T|是所有的新闻文章，tfw,t 是文章T中单词w出现的频率，分母是文章t中出现这个词的数量。

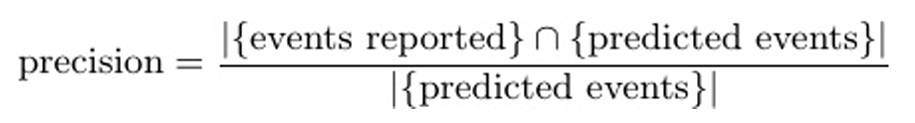


第三个想法：加入时间戳和滞后性。第一篇文章有详细的时间，时间被划分成离散的块或段，以便对某个关键词在不同时间段内的趋势进行分析和比较。相比较第一篇文章来说，第五篇文章的数据是较少的，更加离散的，但是我们至少能有划分为天数的时间戳。此外从第一篇文章可以看到，不同事件在Twitter和eBay上的滞后性不大相同。例如“对于‘服装’，只有45.6%的人具有正滞后值，而对于‘体育’，这一比例为70.14%”。在不同事件中我们也可以横向比较，例如如果我们发现罢工比疫情传播更加迅速，那么我们可以将时间信息也加入编码中，使用MLE，通过观测数据来估计参数 *θ，*更准确地预测事件发生的概率

P\left(\operatorname{ApplyAbs}\left(e v\_{j}(\tau+\Delta), a b s\left(f\_{k}, e n\right)\right), g e n\left(f\_{k}\right)\left(e v\_{i}(\tau)\right)\mid \theta\right)

纵向上我们也可以画出图来看一看各个地区霍乱发生的滞后性，并且分析可能的原因。

第三个想法：使用的是召回率，因为预测从大样本中选取小数量样本，这个时候如果用准确率，可能会得到99%（很多事件没有发生，也预测不发生）,但如果用召回率，则可能是0%



文章在预测死亡人数的时候直接算了召回率和精确度，在算死亡的召回率的时候没有说具体吧用什么方法，但是在文中预测死亡分为三个bin，分别为“少于10人”、“10-100人”和“100人以上”，这是一个三分类问题，或许可以同时算一下Macro-averaging和Micro-averaging，通过算出来的结果找到可能是哪里出现了问题。就像macro是总绩点，micro是每个科目具体的小分。

我觉得在具体的三分类任务中还可以算一下precision和召回率，选择Macro-averaging和Micro-averaging中效果较好的一个作为精确任务的判断标准，和上面对于事件是否发生的二分类任务保持一致，使得文章更加整齐好看。

召回率在判断死亡人数的时候出问题

Macro总绩点，micro每个科目具体的小分

