# 论文阅读报告

# CrystalBall: A Visual Analytic System for Future Event Discovery and Analysis from Social Media Data (对于社交媒体数据进行未来事件识别和分析的可视化分析系统)

**作者：Isaac Cho , Ryan Wesslen , Svitlana Volkova , William Ribarsky , Wenwen Dou**

**论文：<https://wesslen.github.io/assets/documents/papers/crystalball.pdf>**

1. **应用背景**

社交媒体数据可以为世界各地发生的事件提供宝贵的见解，且事件本质上是时间和空间的。现有的文本可视分析系统专注于探测和分析过去和正在进行的事件，很少有人利用社交媒体信息来寻找未来可能发生的事件。在本文中，我们提出了一个交互式可视化分析系统CrystalBall，它可以自动识别和排列来自Twitter的未来事件。CrystalBall集成了新的方法，通过交互式可视化发现事件，从而可以识别未来事件。该方法利用有关的时间，地点，社交网络和消息，整合了七种不同的方法来识别和表征未来事件。同时，可视化界面与计算方法紧密结合以提供未来可能事件的简明摘要。

1. **介绍**

这是一个利用社交媒体数据能够预测未来事件的系统，针对一个特定的事件，如果我们关注的是过去或者正在进行的事件，则这些事件通常被定义为主题和时间，同时伴有人物和位置属性；当我们关注的是将来的事件时，由于数据集的限制，因此这些事件通常被定义为位置和时间，同时伴有主题和社会网络的信息，这里就是根据时间和地点进行的详细分析。

1. **功能介绍**

这篇文章主要是根据twitter数据来预测未来几周可能发生的事件，满足人们的需求。其中数据的收集都是在线进行的，每天刷新以显示未来一段时间内可能发生的最新事件，其工作原理如下:

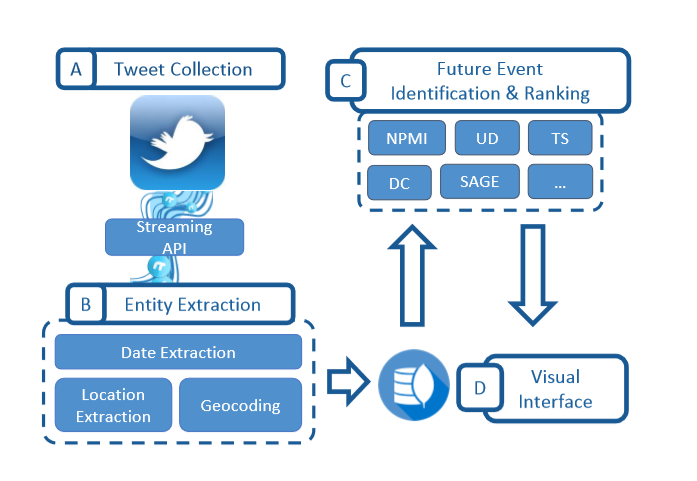


图1.工作原理图

其中，前两个部分分别是Twitter文章的获取和数据预处理，使用的都是常用的自然语言处理的方法进行实体提取。

对于第三部分，即未来事件的识别和排序，他们设计了7个指标，信息如下:

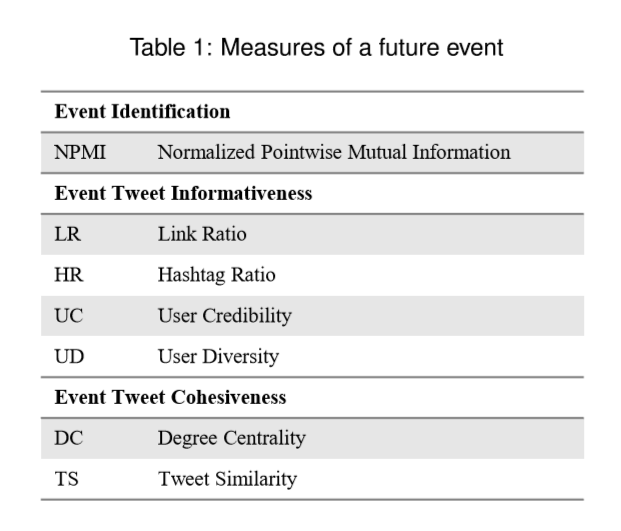


图2.指标信息

1. NPMI识别未来事件。这里考虑了未来事件中的两个重要属性：地点和时间，计算两者的概率分布如下：

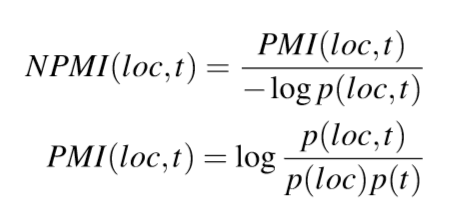


图3.概率分布

1. 对于事件的信息程度的划分，一共考虑了四个维度，包括链接的比例、标签的比例、用户的可信度、用户的多样性。
2. 对于事件的内聚性，设计了两个方面的衡量方式，中心度和Tweet相似性，前者根据用户间的转发和@关系来衡量。
3. 对于事件的排序部分，设计了 一个RanSVM模型，使用监督学习的方式，训练集来自事先从Tweet提取的数据，其中输入为这7个衡量指标的值，输出为对事件的排序等级，具体等级划分如下：

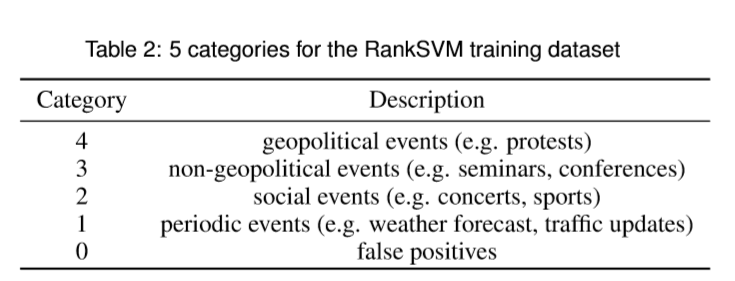


图4.事件的等级划分

对于第四部分，可视化分析系统，主要设计了四个视图来分别表示事件的时间（WHEN）,地点（WHERE）,事件（WHAT）,人物（WHO）,分别对应于下图的A，B，C，D。

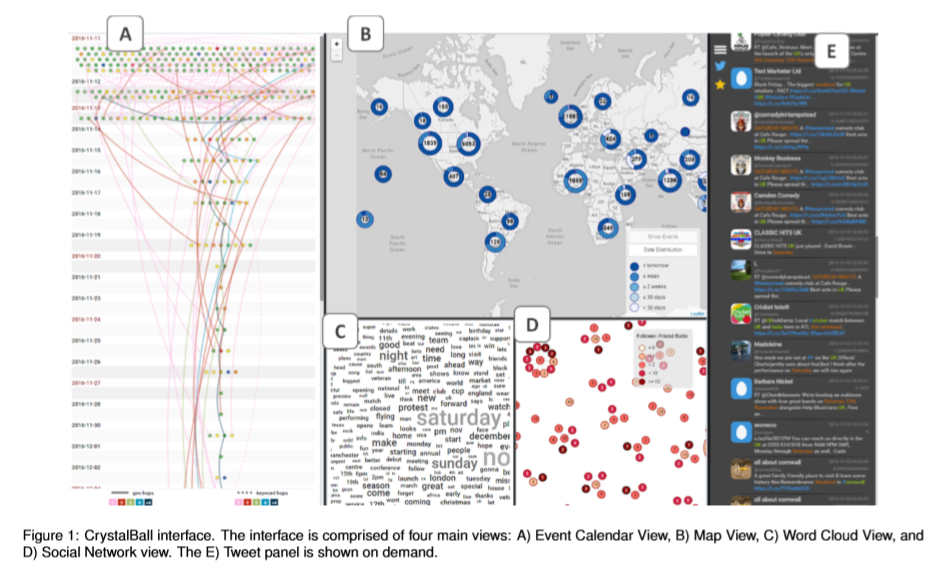


图5.Crystall整体视图

1. 视图A中的每个点表示一个事件，点的颜色表示用户对这个事件的情感。其中实线表示两个事件拥有相同的地点，虚线表示两个事件拥有相同的关键词，用户选择一个事件可以分析各个具体的衡量指标的值。

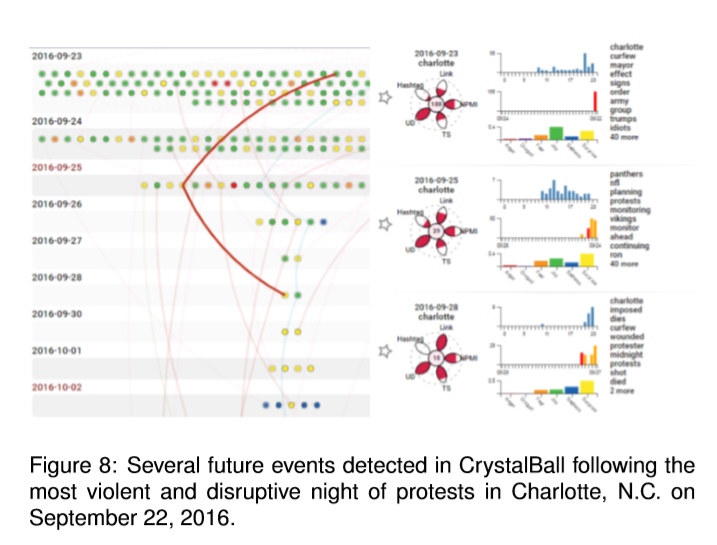


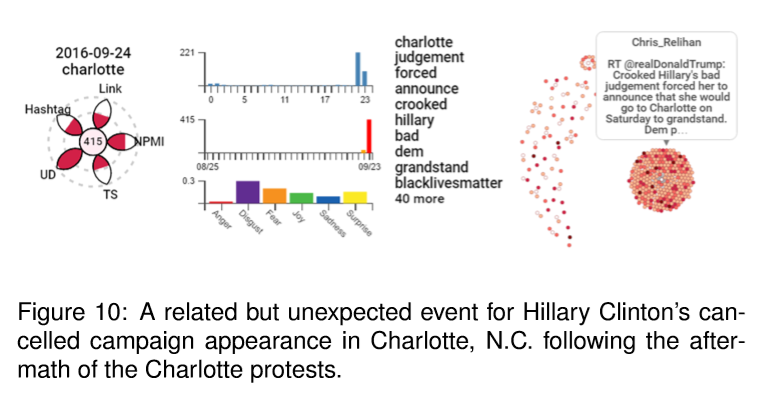
1. 视图B中的圆环的颜色表示各个事件发生的时间距今的远近。
2. 视图C用Word Cloud来展示事件之间的关键词。
3. 视图D显示了用户之间的联系，连线表示转发和@关系

综上所述，这篇论文设计了一个可视分析系统，同时设计了7个维度的信息来对未来事件的识别、排序、描述以及可视化的表示，帮助用户发现未来一段时间内可能发生的事件，提前做好准备工作。

1. **测试数据**

本文中一共介绍了4个例子，这里我们来说明其中的一个。这里介绍的是一件北卡罗来纳州夏洛特市2016年9月抗议活动相关的一周的数据，下图中实线连接的代表了三个有同样的地点的事物，经分析发现，这个时间线中有很多关于抗议的关键词；关注9月24日这一天，可以发现很多人的情绪都转变为恐慌、愤怒。





1. **该系统的优缺点**
2. 优点:通过实际数据进行的预测（可靠的时间和地点）可信度相对较高，能够预测未来可能发生的事件，提醒用户，同时能够做出相应的应对措施，这将大大提高人们的生活防范意识。
3. 缺点:该论文使用 时间-地点 组合进行编码，比较局限；

可能识别到错误的事件信息；

时间、位置的提取算法不是很准确。

以上就是我对该系统的大致体会，尽管能够实现其功能，但是仍有许多不足之处，不仅仅是时间-地点进行组合预测，而应考虑更多因素，像人物，突发事件，以及天气情况等等，数据的多样化能够使结果更加准确。