选题方向

- 社交媒体机器人识别

  通过提供的数据集（其中包含***数个社会热点事件的全部数据及干扰数据***），对于其中的***社交媒体机器人进行判断***，并***根据其行为规律建立识别模型***。

  参考文献
  SBotMiner: Large Scale Search Bot Detection
  提出一种新的大规模从查询日志中检测搜索机器人流量的系统SBotMiner。工作重点是识别和分析以前难以发现的隐蔽的、分布式的、低速率的搜索机器人组。

- ❤️ 核心用户挖掘与传播规模预测

  基于***30条***热门微博的全部***传播数据及参与传播的账号关系***（屏蔽个体身份识别信息后），***找出各条微博中***的***核心传播者***，并***依据***核心传播者的***行为集合***，***建立模型***，***预测单条微博的传播规模***。

  1.核心用户识别(reading)
  2.每条微博都会存在核心传播者(<=30)
  3.训练模型

  - https://dl.acm.org/conference/wsdm
    ACM国际Web搜索和数据挖掘会议论文集
  - Measuring User Influence in Twitter: The Million Follower Fallacy
    百万追随者谬论
    - power-law characteristic：幂律特性

      用户的影响程度可能因数量级的不同而不同：最具影响力的文章被转发或提及的次数超过了大多数用户。

    - 核心用户判断
      1.单凭indegree(粉丝数)很少揭示用户的影响；
      2.retweets由tweet的内容价值驱动，而mention由用户的名称价值驱动；
      3.最有影响力的用户对各种话题都有很大的影响力。

      注意⚠️
      1.用户发表的推文数量与其关注的人数是无用的，因为他们分别将机器人和垃圾邮件定义为最有影响力的；
      2.流行的关键词通常会在一定时间后受到垃圾邮件的影响。

    - 推演模型(理论上)
      1.Twitter的顶级用户拥有不成比例的影响力，这是由power-law distribution所证明的；
      2.主流新闻机构在不同的话题上产生高水平的转发。名人更容易被转发，这是因为他们的name value，而不是content value；
      3。影响不是自发或偶然获得的，而是通过concern efford。为了获得和保持影响力，用户需要保持极大的个人参与。

      ⚠️ 数据分析以转发和提及为重点

- Learning Influence Probabilities In Social Networks
  WSDM '10：第三届ACM国际Web搜索和数据挖掘国际会议论文集 2010年2月

  一个亮点是，除了预测用户是否会执行某个操作外，我们的算法对具有高影响力得分的用户的预测往往具有较高的精度。

- 社交媒体实时热点发现

  通过***含有热点事件及干扰信息***的数据集，***根据***微博***发布时间模拟实时数据流***及时***发现***社交媒体平台上的***热点事件传播规律***，并通过这些热点事件的***共性特征建立模型***，用以***发现***社交媒体上的***潜在热点***。

- 基于新媒体传播数据的地域舆论环境感知

  通过对某一地域阶段内全部新媒体传播数据（含***新闻媒体数据、社交媒体数据***等）进行挖掘，***建立用于感知此地域舆论环境的评价指数模型***，包括但不限于“居民幸福感指数”、“营商环境健康指数”、“城市形象指数”、“焦虑感指数的时间空间社群分布及演化”。

- 突发事件发展推演模型

  基于多个热点事件的***发生、发展传播特征***，找出事件***演化规律***，形成***推演模型***。

- 2019竞赛作品：
  给定数据集 -> 自行决定探究问题方向 -> 数据分析 -> 得出结论

  - 2020竞赛作品：
    给定数据集 -> 给定需要解决问题方向 -> 要求建立相关对应模型 -> 数据分析与建模 -> 得出结论