哈爾濱工業大學
社会计算与信息检索研究中心

中文微博事件情感分布的原因分析

作　者：李泽魁，赵妍妍，秦兵，刘挺
演讲者：段俊文
单　位：社会计算与信息检索研究中心
院　校：哈尔滨工业大学

中文微博事件情感分布的原因分析
—— 大纲

- 一些概念的解释
- 当前研究的挑战
- 算法设计
- 实验部分
- 总结

中文微博事件情感分布的原因分析
—— 大纲

- 一些概念的解释
- 当前研究的挑战
- 算法设计
- 实验部分
- 总结

中文微博事件情感分布的原因分析
——中文微博的情感

- 微博(Micro-blog)
  - 推特 (twitter.com)
  - 新浪微博(weibo.com)
  - ……

- 中文微博
  - 中国网民获取、传播及分享身边的新鲜事的媒介
  - 日均产生千万级的微博数据
- 中文微博的情感在本文中的定义
  - 喜悦、愤怒、悲伤、恐惧、惊奇

中文微博事件情感分布的原因分析
——中文微博事件

- 微博事件
  - 负面事件
    - 灾难恐怖事件
    - 娱乐八卦事件
    - ……
  - 正面事件
    - 正面新闻事件
    - 网民热议事件
    - ……

天津港爆炸事件

讨论人数：2691356

王林被警方带走

讨论人数：25392

9.3抗战胜利阅兵

讨论人数：127197

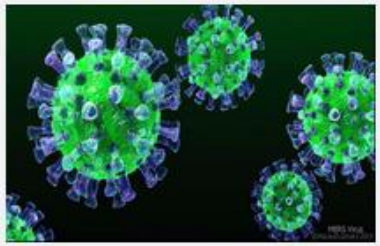
王思聪范冰冰骂战

讨论人数：42252

更多事件举例详见 [Link]

中文微博事件情感分布的原因分析
——事件情感分布

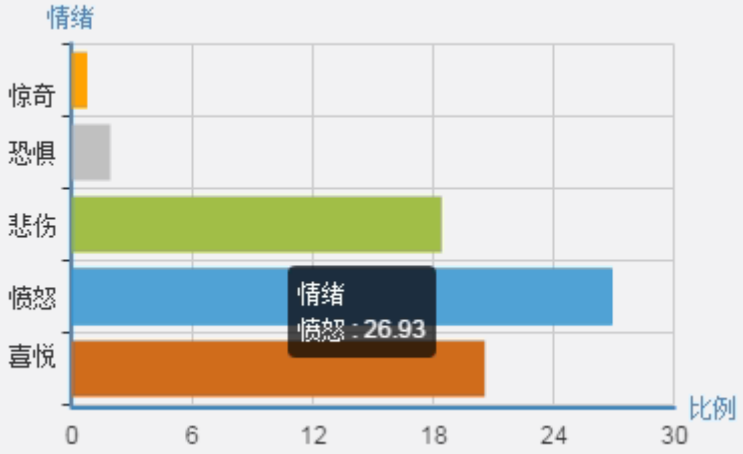
- 微博的情感
  - 喜悦、愤怒、悲伤、恐惧、惊奇
- 微博事件的情感分布



MERS入侵广东

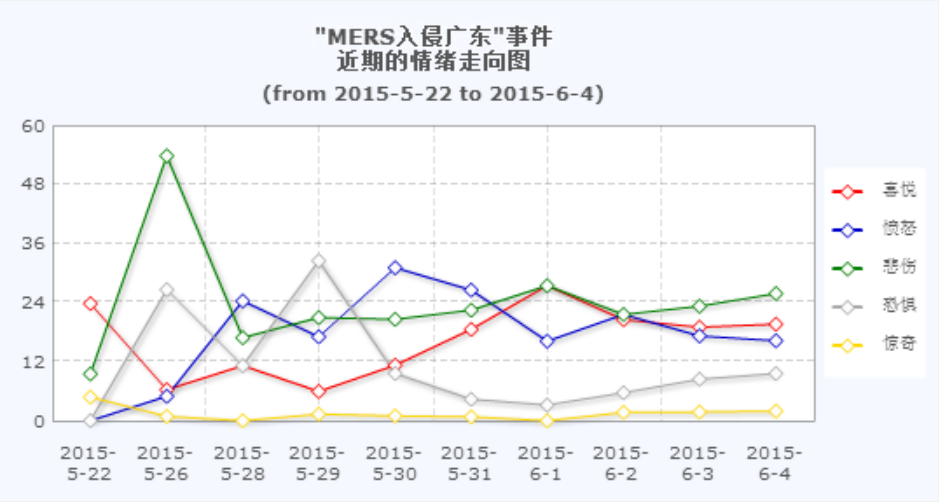
讨论人数：237520

事件简介：广东确诊一例从韩国输入的中东呼吸综合征病例，该病例于26日乘坐韩亚航空OZ723航班到香港，后下午三时从机场搭大巴至深圳沙头角，再于下午4时46分乘大巴至惠州。现呼吁：曾与其同乘者请主动与省疾病预防控制中心联系。



中文微博事件情感分布的原因分析
——情感分布原因

- 研究的意义 (一)
  - 挖掘事件发展过程中情感变化的依据



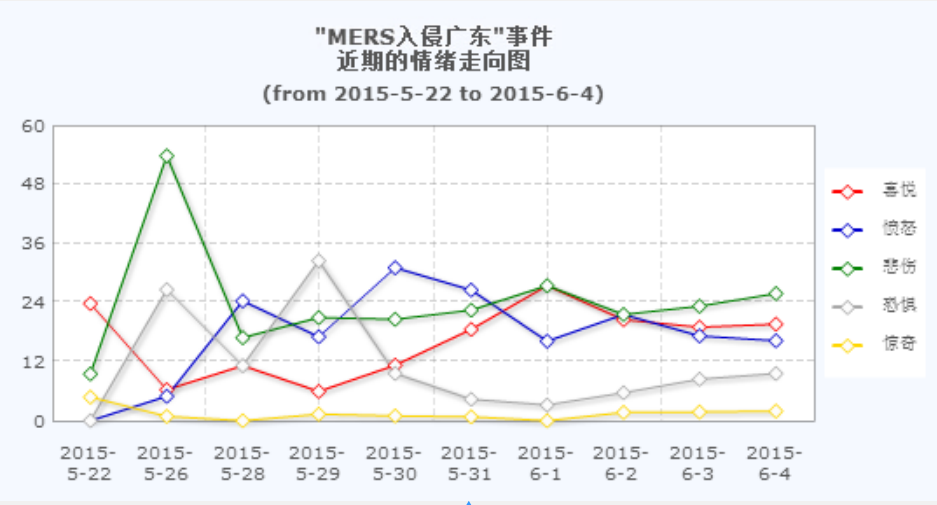
日期	事件相关新闻
5月22日	MERS病毒在韩国出现首例
5月26日	病毒携带者在广东确诊
5月29日	隔离治疗的患者病情加重
5月30日	韩方的监管漏洞及技术乏力
6月01日	惠州医院护士抽签上岗

“微博情绪指数”系统的自动分析结果 ⟷ 对事件相关新闻的人工分析结果

是否有关联?

中文微博事件情感分布的原因分析
——情感分布原因

- 研究的意义 (一)
  - 挖掘事件发展过程中情感变化的依据



日期	事件相关新闻	情绪
5月22日	MERS病毒在韩国出现首例	喜悦
5月26日	病毒携带者在广东确诊	悲伤
5月29日	隔离治疗的患者病情加重	恐惧
5月30日	韩方的监管漏洞及技术乏力	愤怒
6月01日	惠州医院护士抽签上岗	喜悦*

自动分析结果与人工分析结果相符

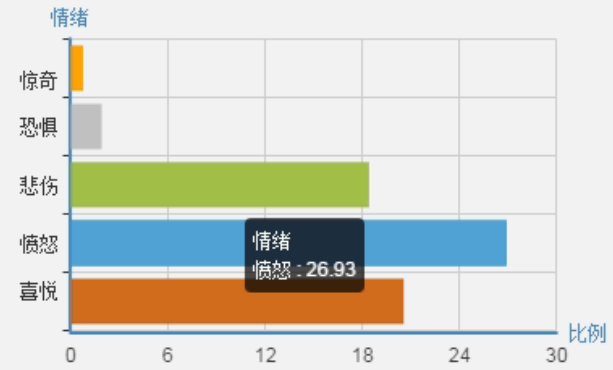
*注:“喜悦”在上表中含义不同

中文微博事件情感分布的原因分析
——情感分布原因

- 研究的意义 (二)
  - 挖掘情感分布对应的话题

MERS入侵广东

讨论人数：237520

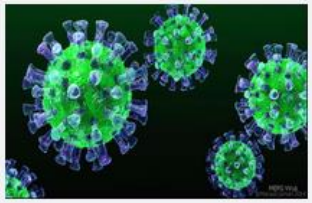


影响分布的话题有哪些?

ID	事件相关的话题
1	强行出境MERS患者令韩国人蒙羞
2	中国在MERS上小题大做
3	广东未婚医护人员抽签上岗
4	韩国卫生部长就MERS问题向中国感谢
5	中国首例MERS患者病情加重
6	患中东呼吸综合征的韩国人为啥能出国
7	急寻与韩国MERS患者同行旅客
8	MERS病死率高于SARS无药物治疗方法
9	韩国MERS患者病中来华因太忙
…	……

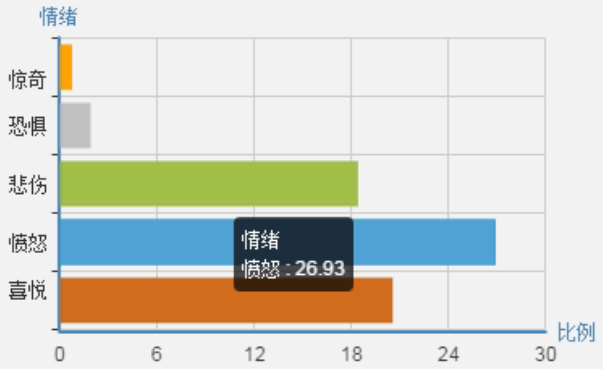
中文微博事件情感分布的原因分析
——情感分布原因

- 研究的意义 (二)
  - 挖掘情感分布对应的话题

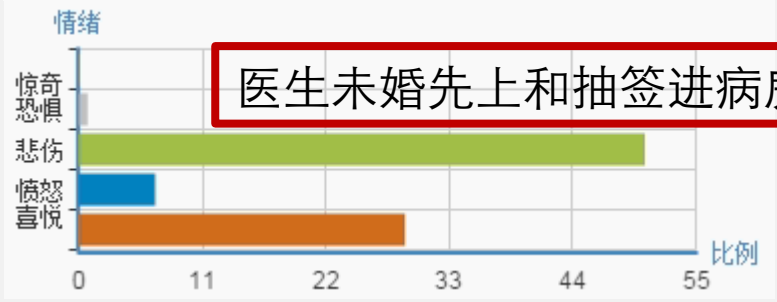


MERS入侵广东

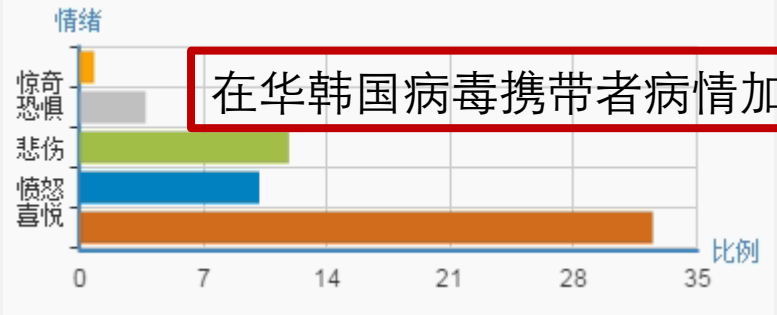
讨论人数：237520



韩国卫生部门监管乏力



医生未婚先上和抽签进病房



在华韩国病毒携带者病情加重

中文微博事件情感分布的原因分析
—— 大纲

- 一些概念的解释
- 当前研究的挑战
- 算法设计
- 实验部分
- 总结

中文微博事件情感分布的原因分析

- 存在的挑战
  - 国内相关研究工作少，实验数据匮乏
  - 对微博话题的分析没有一套完备的算法
  - 本算法依赖一套成熟的中文微博情感分析算法

中文微博事件情感分布的原因分析

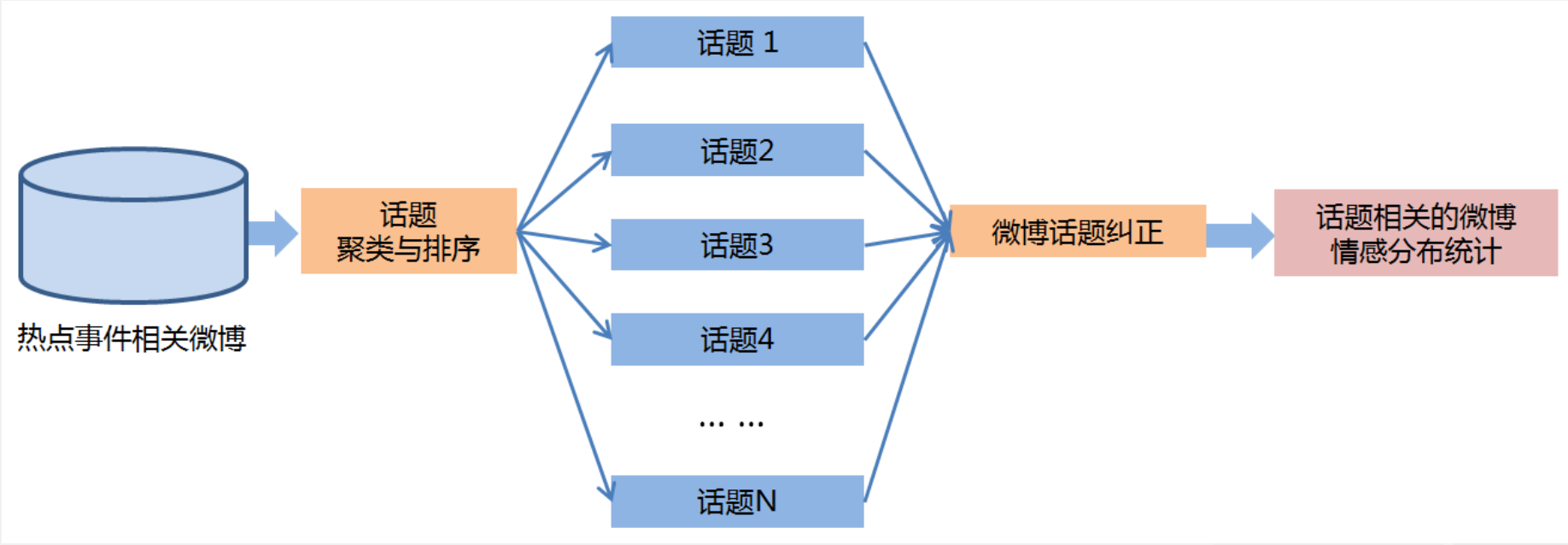
- 存在的挑战
  - 国内相关研究工作少，实验数据匮乏
    - 本文收集并标注了一套任务相关的实验语料
  - 对微博话题的分析没有一套完备的算法
    - 无监督学习的层次聚类排序方法
    - 半监督学习的微博话题纠正算法
  - 本算法依赖一套成熟的中文微博情感分析算法
    - 使用了基于特征抽取的机器学习情感分类算法

中文微博事件情感分布的原因分析
—— 大纲

- 一些概念的解释
- 当前研究的挑战
- 算法设计
- 实验部分
- 总结

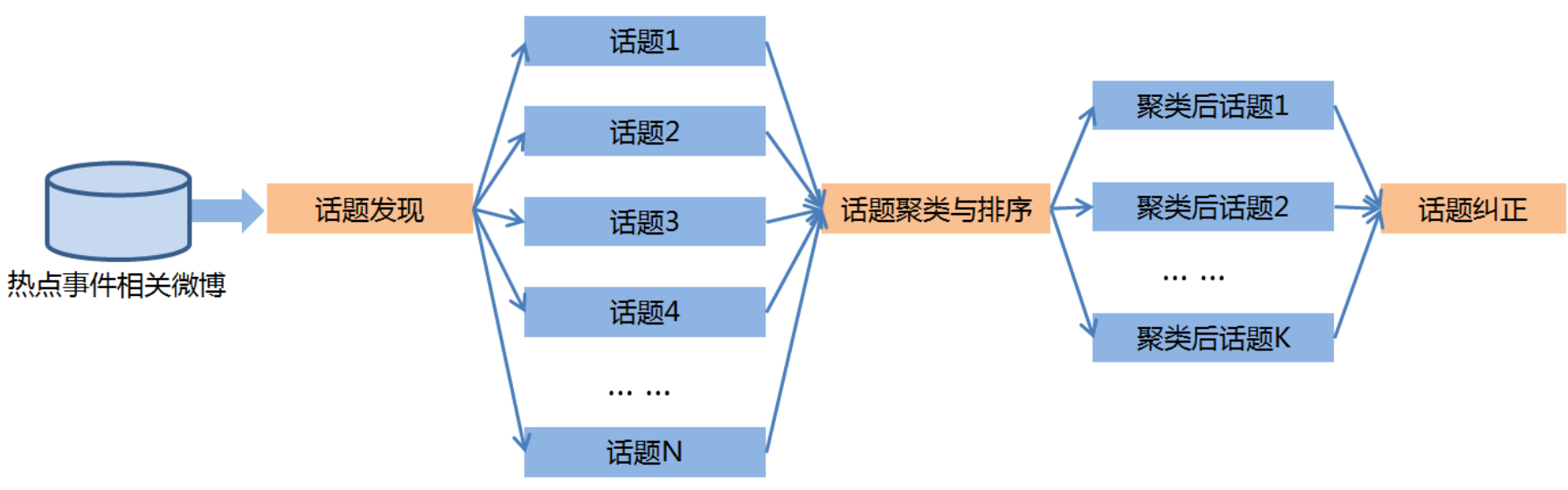
中文微博事件情感分布的原因分析

- 总体流程



中文微博事件情感分布的原因分析
—— 事件话题挖掘方法对比

- 相关研究
  - 微博的转发率、词频等信息对词条排序 (Zhao 2011)
  - 微博的多样性、唯一性、突发性信息对话题进行建模 (Das 2014)
- 我们的方法
  - 对微博的Hashtag信息进行发现、聚类及排序



中文微博事件情感分布的原因分析
—— 话题发现

- 微博的话题
  - 本文使用微博的#Hashtag#信息作为话题信息
  - 实验证明，事件相关的微博在Hashtag的多样性上是有保证的
- 以“庆安枪击案”为例
  - 爬取微博 172,644 句
  - 收集Hashtag 344 个
  - Hashtag 主题分布众多
    - 从“事件主人公徐纯合”到“开枪民警李乐斌”
    - 从“支持警察开枪”到“同情受害人”
    - 从“枪支使用是否合理”到“庆安官场贪污腐败”
    - ……

中文微博事件情感分布的原因分析
—— 话题聚类

- 语料中存在相似的话题
  - #呼吁公布庆安车站枪案现场视频# vs #公开被击毙视频#
  - #庆安县副县长董国生被停职# vs #副县长麻烦来了#
  - ……
- 话题的聚类
  - 对部分相似的话题进行合并
- 聚类的方法
  - 层次聚类 (Hierarchical Cluster) 算法

中文微博事件情感分布的原因分析
—— 话题聚类

- 聚类的方法
  - 层次聚类 (Hierarchical Cluster) 算法
- 聚类距离计算公式
  - 基于微博词频/逆文档频率（TF/IDF）的相似度算法

$$Similarity_{TFIDF}(S_A, S_B) = cosine(TFIDF(S_A),\ TFIDF(S_B))$$

  - 基于微博Hashtag字符串相似度算法

$$Similarity_{Hashtag}(H_A, H_B) = \frac{Length(LCS(H_A, H_B))}{min(Length(H_A), Length(H_B))} + \left(1 - \frac{Edit\ Distance(H_A, H_B)}{max(Length(H_A), Length(H_B))}\right)$$

中文微博事件情感分布的原因分析
—— 话题排序

- 层次聚类的结果
  - U个聚类话题簇
- 根据热度对聚类结果排序
  - 根据微博数目排序

$$Ranking\ Score(\text{topic}) = \text{topic}_{\text{weibonumber}}$$

  - 根据微博数目与簇内话题数的加权关系排序

$$Ranking\ Score(\text{topic}) = log(\text{topic}_{\text{weibonumber}}) \cdot \text{topic}_{\text{num}}$$

中文微博事件情感分布的原因分析
—— 话题纠正

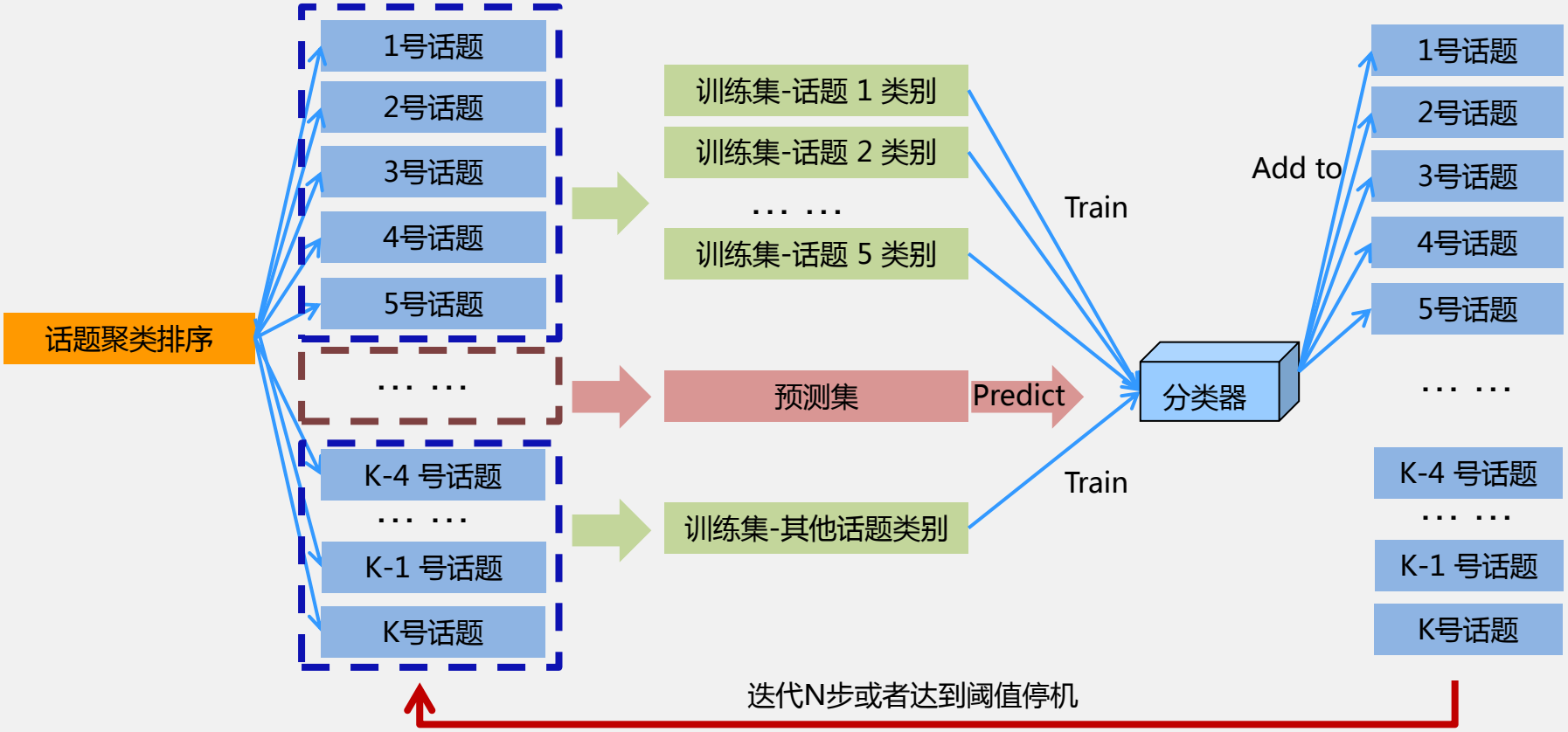
- 话题发现、聚类和排序的结果
  - U个话题排序结果
- 是否聚类排序结果足够接近真实分布?

微博内容	内容指向	正确Hashtag
#火车站里的一声枪响# 让当地的贪官都浮出水面!	当地政府（贬义）	#庆安官场贪腐#
#视频四大疑点# 支持开枪民警!	开枪民警（褒义）	#决不能暴力袭警#

- 微博话题纠正任务
  - 目标是将事件相关的微博划分到正确的主题下

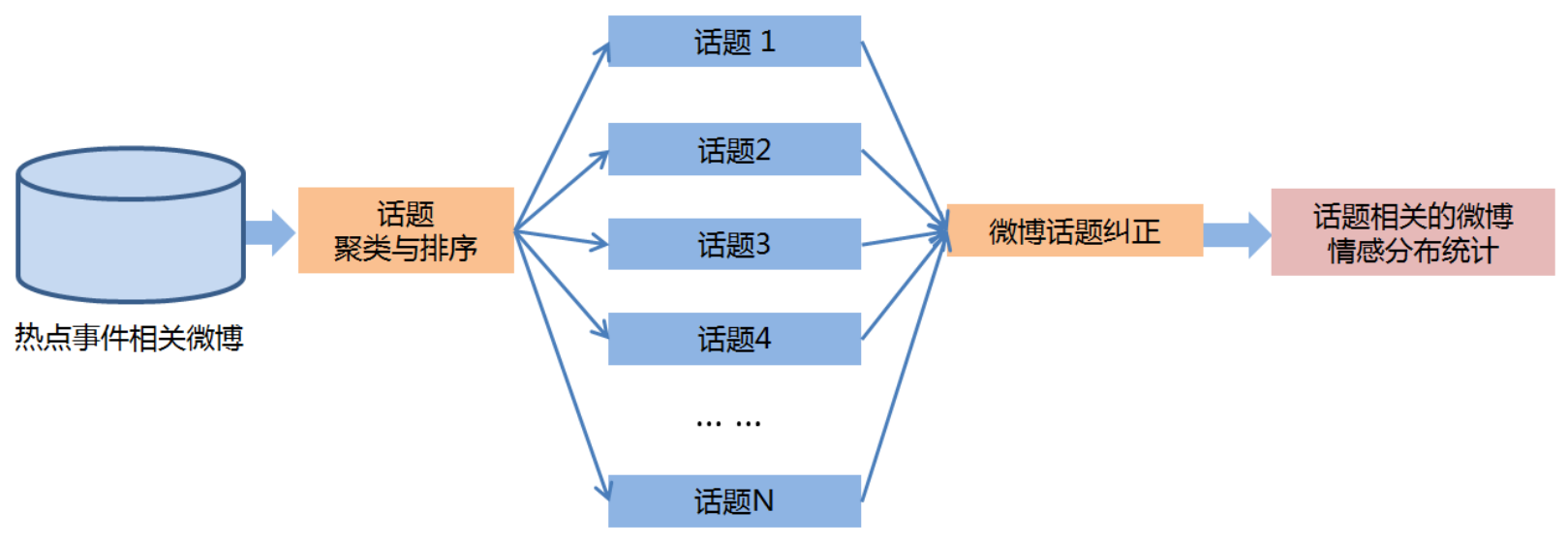
中文微博事件情感分布的原因分析
—— 话题纠正

- 微博话题纠正任务
  - 目标是将事件相关的微博划分到正确的主题下
  - 基于半监督学习的微博话题纠正算法



中文微博事件情感分布的原因分析

- 回顾总体流程



中文微博事件情感分布的原因分析
—— 大纲

- 一些概念的解释
- 当前研究的挑战
- 算法设计
- 实验部分
- 总结

中文微博事件情感分布的原因分析
—— 实验数据

- 微博数据
  - 2015年5月~7月间发生的微博热点事件相关微博
  - 对单个事件的最热门的5个的话题进行人工标注

事件	微博数目
毕福剑不雅视频事件	121,204
云南女导游辱骂游客	18,338
四川女司机被当街暴打	324,775
……	
贵州自尽男童	30,848

事件	ID	代表子话题
庆安枪击事件	1	请央视公布完整监控视频
	2	开枪民警属正当履职
	3	庆安官场全面崩塌
	4	真相别总靠“倒逼”
	5	开枪民警李乐斌偿命

中文微博事件情感分布的原因分析
——对比实验 1

- 话题聚类距离算法对比实验
  - 基于微博TFIDF相似度计算方法
  - 基于Hashtag字符串相似度计算方法

指标\距离公式	Similarity_TFIDF	Similarity_Hashtag
准确率@5（%）	53.3	78.7

- 实验结果
  - 使用TFIDF计算相似度效果不佳
  - 单个事件相关的微博所用词语的重合度比较大，例如“庆安枪击案”相关微博中“视频”、“开枪”等词汇频繁出现

中文微博事件情感分布的原因分析
——对比实验 2

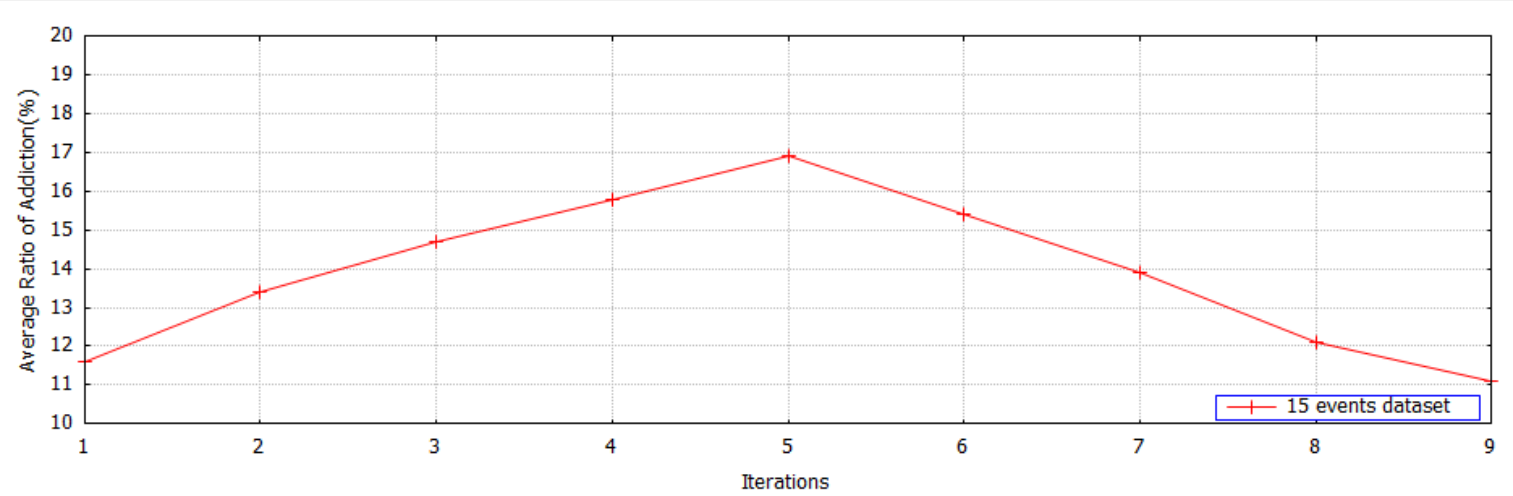
- 话题聚类结果排序算法对比实验
  - 根据微博数目排序 (1)
  - 根据微博数目与聚类结果话题数的加权关系排序 (2)

指标\排序公式	排序方法 1	排序方法 2
准确率@5（%）	66.7	78.7

- 实验结果
  - 仅按照聚类后簇对应的微博数目作为排序指标效果不佳
  - 簇内聚合的话题数目也在很大程度上代表了该话题的热议程度，即话题聚类结果所含话题越多，其热议程度越高

中文微博事件情感分布的原因分析
——对比实验 3

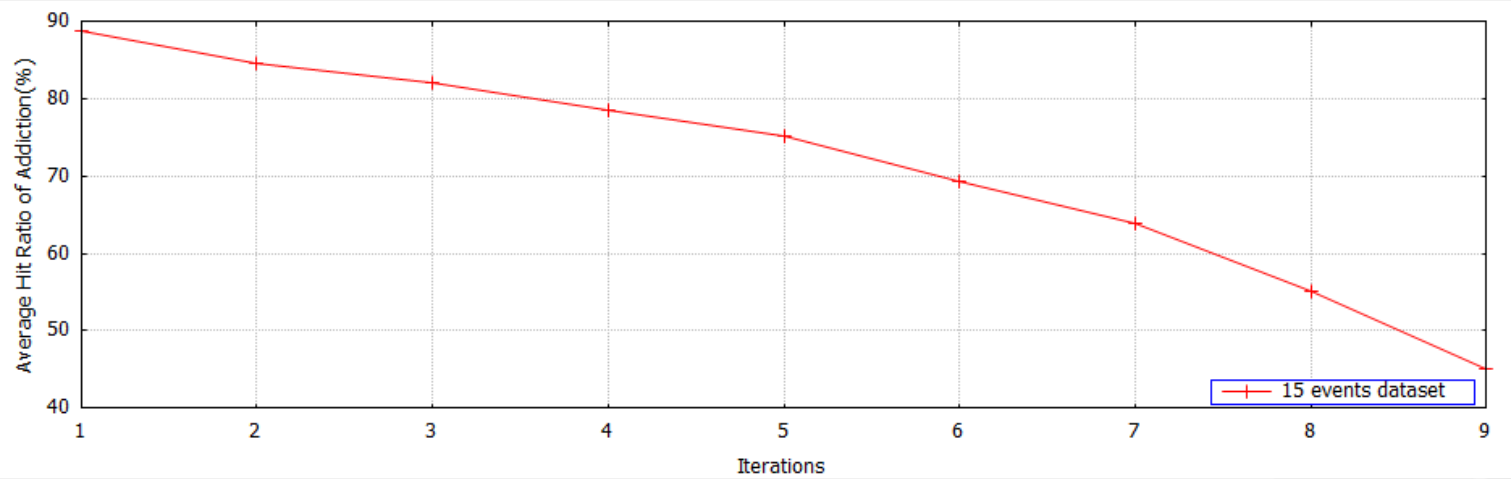
- 微博事件话题纠正实验
  - 随着迭代次数的增加，话题种子集的微博数目平均添加率先增后减



  - 迭代初期，随着分类模型追加的训练微博数量增多，使模型的特征增多，从而可以预测更多的样本
  - 后期由于所剩预测样本数目不足，导致添加率降低

中文微博事件情感分布的原因分析
——对比实验 3

- 微博事件话题纠正实验
  - 随着迭代次数的增加，话题种子集追加微博的平均命中率(即话题补充微博的质量)呈下降趋势



  - 随着模型迭代更新，一方面训练语料不断在增加，另一方面训练噪声也在不断累加

中文微博事件情感分布的原因分析
—— 大纲

- 一些概念的解释
- 当前研究的挑战
- 算法设计
- 实验部分
- 总结

中文微博事件情感分布的原因分析 —— 总结

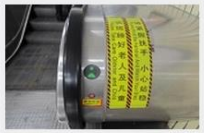
- 国内相关研究工作少，实验数据匮乏
  - 本文收集并标注了一套任务相关的实验语料
  - 15个微博事件，分别标注各自的代表话题
- 对微博话题的分析没有一套完备的算法
  - 无监督学习的层次聚类排序方法 (使用Hashtag相似度计算距离)
  - 半监督学习的微博话题纠正算法 (迭代的算法对话题微博进行追加)
- 实验结果证明本方法的有效性

中文微博事件情感分布的原因分析
—— 更多分析结果

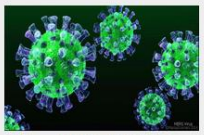
- 更多分析结果链接 [LINK]
- 欢迎各位老师、同学提出宝贵意见！


9.3抗战胜利阅兵

天津港爆炸事件


荆州电梯吞人事件


深航纵火事件


MERS入侵广东


故宫女模裸照

携程被黑


吴镇宇发飙


恶搞花木兰道歉

优衣库视频事件


王林被警方带走


刘翔离婚


王思聪范冰冰骂战


庆安枪击事件


男子暴打扫地小男孩


四川女司机被当街暴打


宝马撞散马自达


粤赣高速匝道坍塌


邓超出轨


长江游轮倾覆


云南女导游辱骂游客


李一男创业


毕福剑不雅视频事件


贵州自尽男童


微博炫腹大赛


肃宁县特大枪击案


边策吸毒坠楼身亡

哈爾濱工業大學
社会计算与信息检索研究中心

哈爾濱工業大學
社会计算与信息检索研究中心

Thanks Q&A