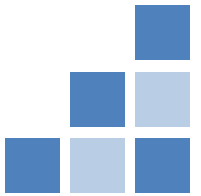
北京大学
PEKING UNIVERSITY

高校BBS与微博用户社交行为特征分析

报告人　赖清楠

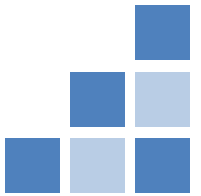
目录

Contents



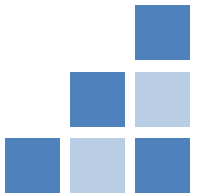
一、背景

- 高校BBS已成为校园文化不可或缺的一部分，是大学生交流、娱乐与生活的重要网络平台。
- 微博在信息传播的速度和广度方面已远远超过传统媒体。
- 实现两个平台信息的互联共享很有必要，出现了许多高校BBS官方微博。
- 以北京大学未名BBS官方微博为研究对象。



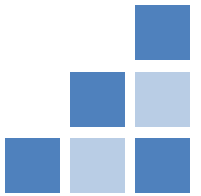
目录

Contents



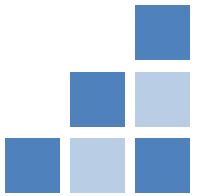
二、微博信息抓取与编辑

- 数据采集
  - 使用新浪微博提供的API接口，采集的数据包括微博和评论内容。
- 预处理
  - 去除URL。
  - 去除包含的转发内容。
  - 去除@符号及其后面的用户名。
- 分词
  - 经过预处理后的数据进行分词。
- 微博编辑
  - 将微博内容进行整合以便分类查看。

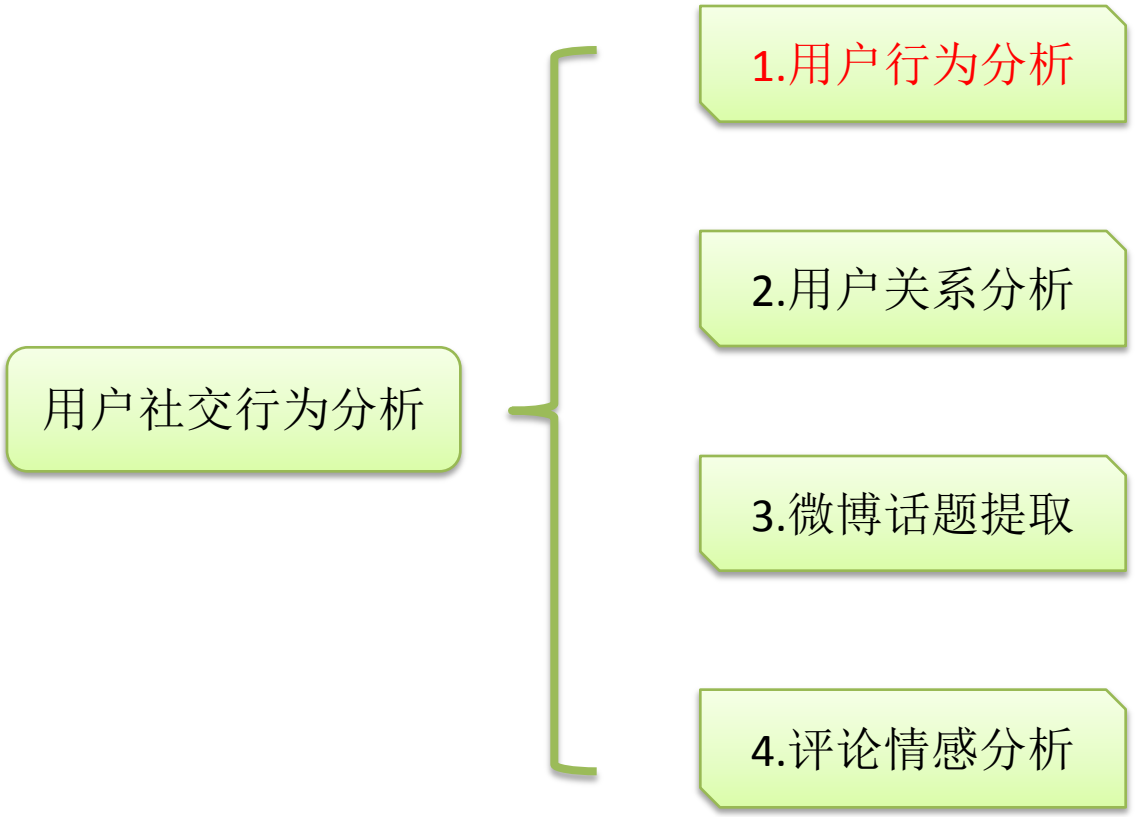


目录

Contents



三、用户社交行为分析



1、用户行为分析

- 活跃度
  - 微博活跃度：发表微博数量占某段时间发表微博总量的比，微博活跃度较高，说明用户关注的事情较多。
  - 评论活跃度：收到评论数量占某段时间收到评论总量的比，评论活跃度较高，说明其他用户对用户发表的微博比较感兴趣。

$$\text{活跃度：} A_t = \frac{sum_t}{sum}$$

A_t – 表示活跃度；sum_t – 表示子时间段t内所发微博数量/所接收到评论的数量；

sum – 表示某段时间内所发微博总数/所接收到评论的总数

表1 微博和评论数量月统计

月份	4	5	6	7	8	9	总计
微博总数	252	182	137	102	101	67	841
评论总数	2285	1367	883	548	726	718	6527

1、用户行为分析

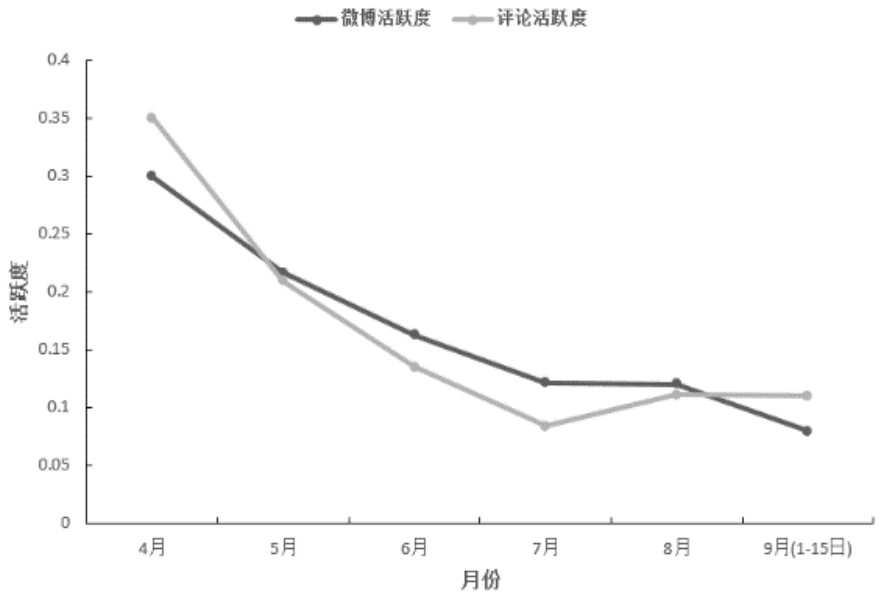
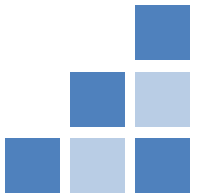


图1 4-9月份活跃度统计

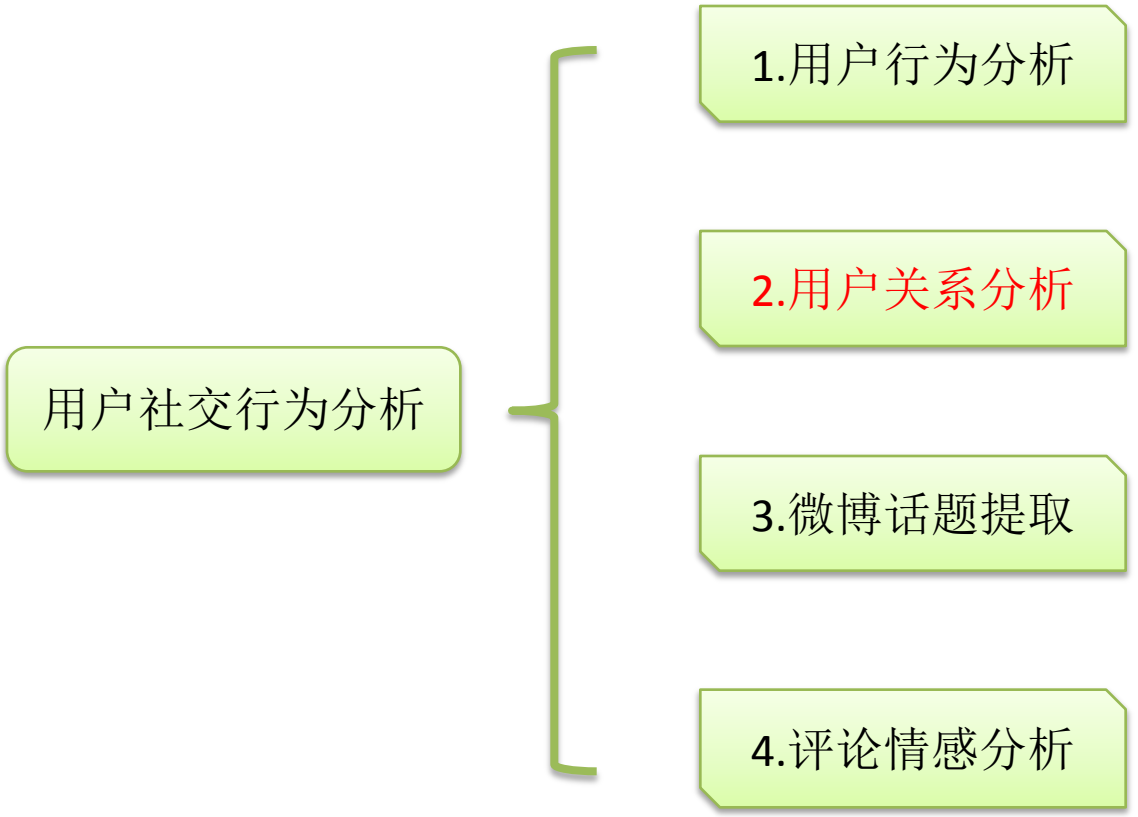
4月到7月活跃度递减，4月份刚开学，活跃度较高，7月份学期已经结束，活跃度较低，9月份随着新学期的开始，活跃度逐渐升高，评论活跃度与微博活跃度的变化趋势基本是一致的。

1、用户行为分析

通过对活跃度的计算，可以衡量北大未名BBS微博用户在微博上的行为，可以知道某段时间热点事件多少及网友对事件的关注程度。



三、用户社交行为分析



2、用户关系分析

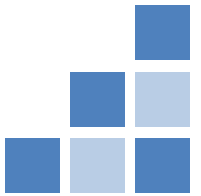
- 分析用户之间的关系，找出用户之间的关注程度以及潜在的好友关系。
- 微博用户之间仅有关注和未关注的关系，关注的人未必就会是你的好友。
- 紧密度和亲密度
  - 紧密度：用来表示两个用户之间的关注程度，通过用户对微博的评论的次数来计算。
  - 亲密度：用来表示两个用户之间的好友关系，通过用户之间的交流次数来计算。

紧密度：$T_{21} = \frac{r_{12}}{r_1}$

T_{21} —用户a_2对a_1的紧密度；
r_{12} —用户a_1与a_2之间的关系个数；
r_1 — 用户a_1与其他用户的关系总数。

亲密度：$I_{21}=m_1+m_2$

I_{21} —用户a_2对a_1的亲密度；
m_1 —用户a_2对a_1微博的评论数；
m_2 — 用户a_1对a_2评论的回复数。



2、用户关系分析

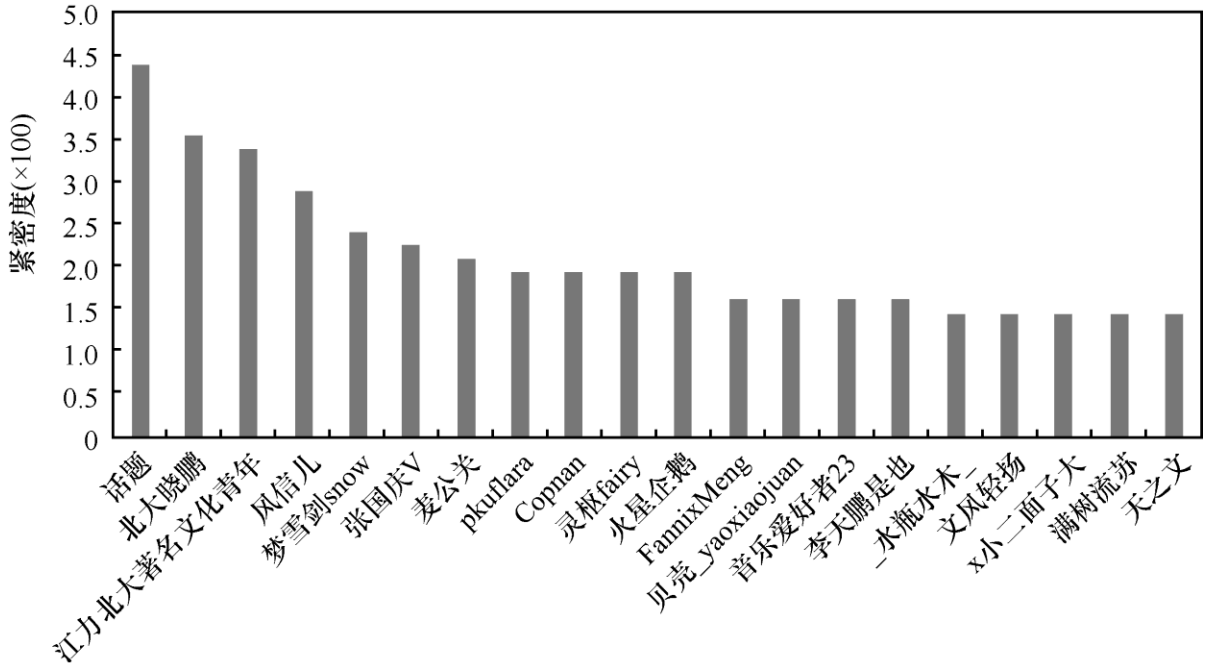


图2 用户紧密度

紧密度越大，则表示该用户对北大未名BBS关注度越高。

2、用户关系分析

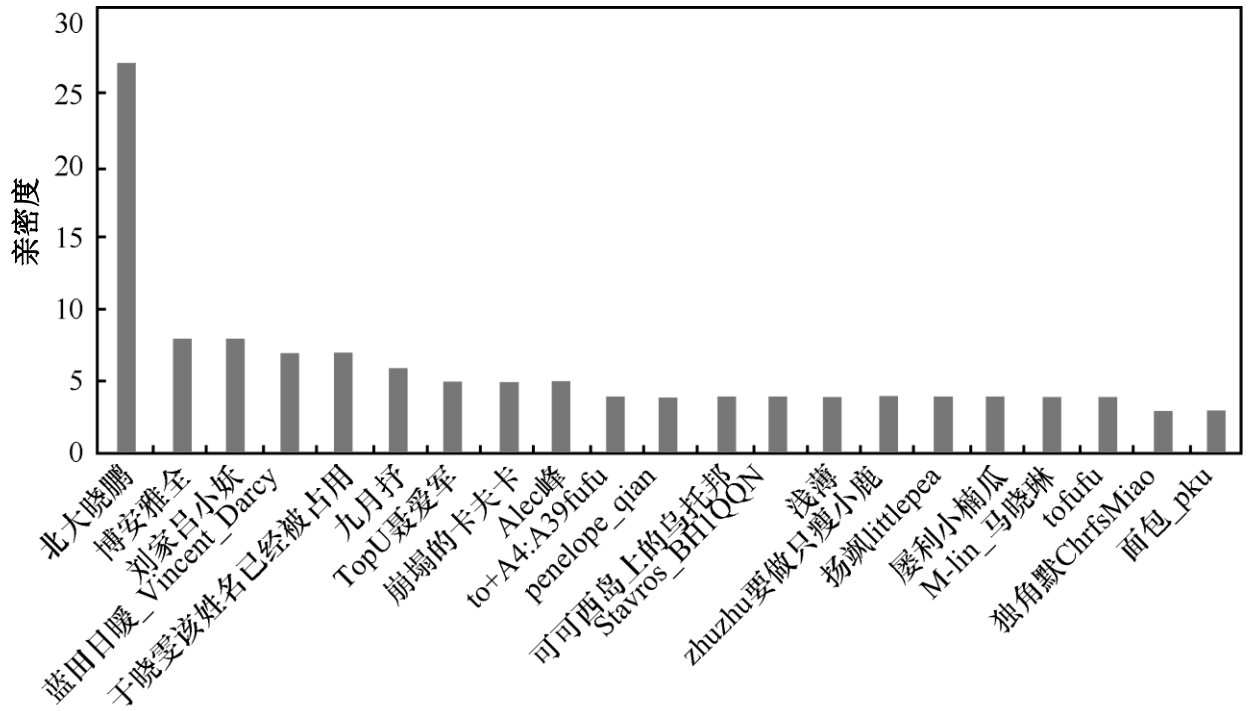


图3 用户亲密度

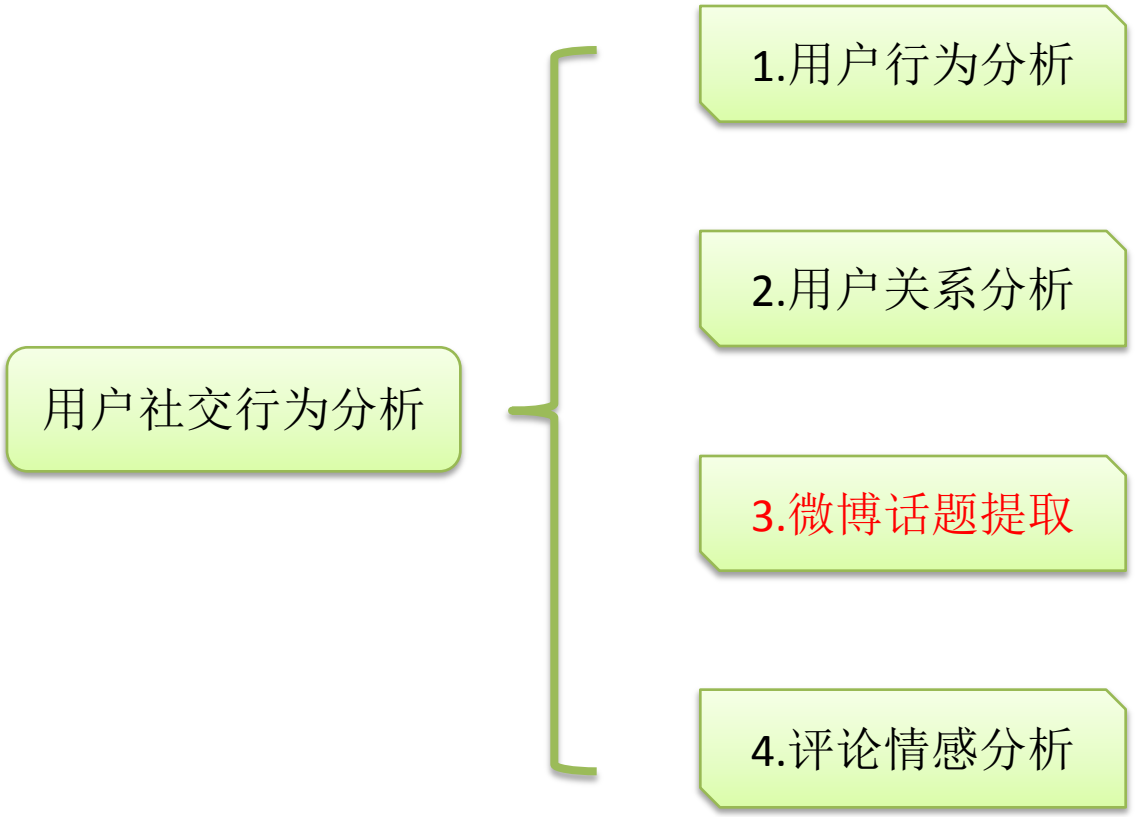
亲密度越大，则与北大未名BBS是好友的可能性越大

2、用户关系分析

使用紧密度和亲密度能很好的描述用户之间的关注程度及潜在好友关系，上述结果表明，与北大未名BBS是好友关系的用户对北大未名BBS的关注度不一定就高，紧密度与亲密度两者之间没有必然的联系。

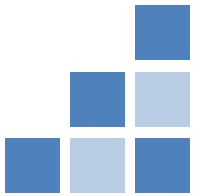
综合BBS微博关注程度以及好友关系分析，用户名中带有典型的北大标识，比如“pku”、“北大”、“北京大学”，这些用户的数量约占14%，由此可推定北大未名BBS微博受到了北大师生较为广泛的关注，也为北大师生提供了一个很好的交流平台。

三、用户社交行为分析



3、微博话题提取

- 微博话题的提取目的是简明直接的了解微博内容，以便找出热点事件。
- 采用基于特殊标点符号的微博话题提取方法。
  - #XXX#之间以及一些其他特殊符号如《》、“”、【】之间的内容提取出作为微博的话题。
- 实验验证，采用特殊标点符号进行话题提取能达到较好的效果。

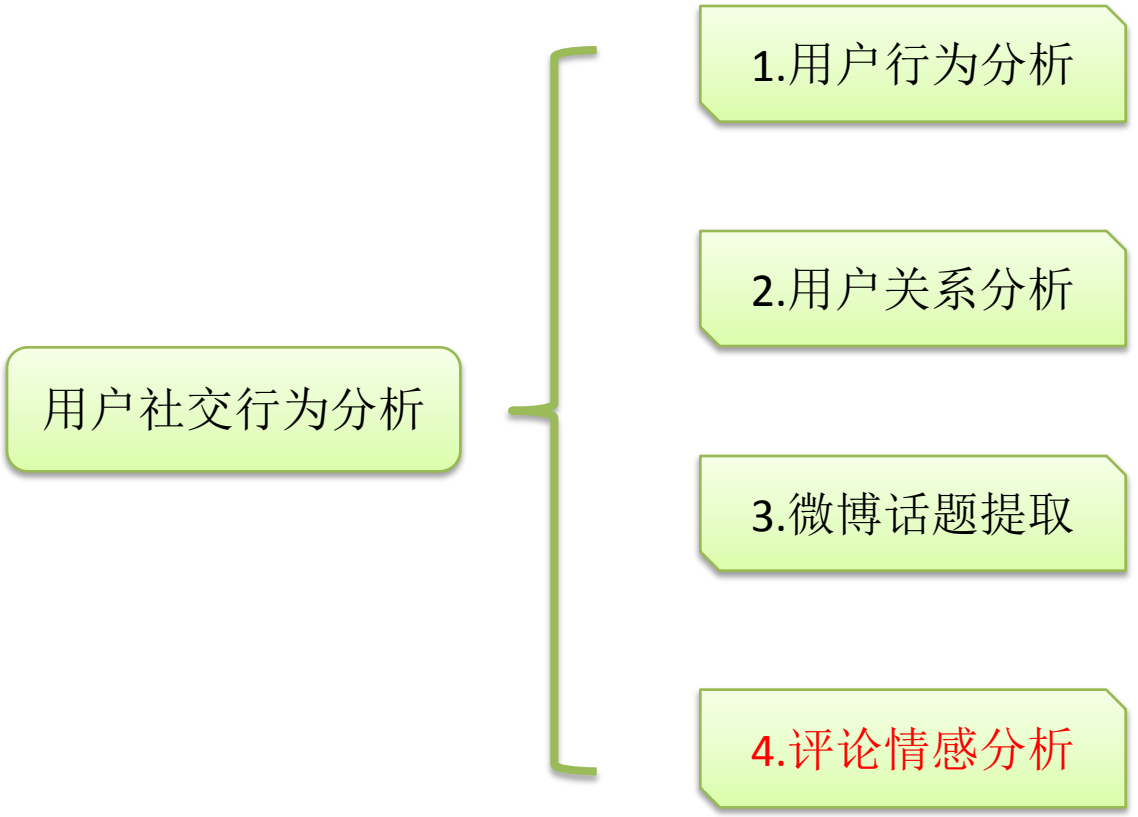


3、微博话题提取

表2 TOP10热点事件

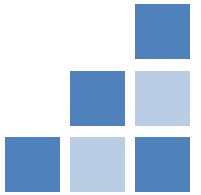
排名	时间	评论数	热点事件
1	2013.4.1	1012	炫彩燕园明信片微博首秀
2	2013.5.4	143	115周年校庆、未名BBS十大贴《师弟，听师姐和你讲》
3	2013.5.3	139	北大115周年校友返校
4	2013.4.22	116	周杰伦百讲录制央视节目心系雅安芦山正能量传递活动
5	2013.9.4	116	燕园新变化之二教艺术品
6	2013.8.31	104	迎新工作
7	2013.9.6	90	免费发放燕园元素卡贴
8	2013.7.9	84	北京大学2013年毕业典礼
9	2013.9.5	80	新生明信片闪亮登场
10	2013.7.8	79	“北大未名”杂志创刊号明天首发

三、用户社交行为分析



4、评论情感分析

- 评论一般都是带着评论人的情感，对评论进行情感分析可以知道网友对用户所发微博的态度，给用户一个反馈。
- 方法
  - 标注评论的情感极性
  - 否定词的处理
  - 基于词典的情感分析
    - Xsimilarity中文情感词典和自己构建的网络词典
  - 基于表情符号的情感分析



4、评论情感分析

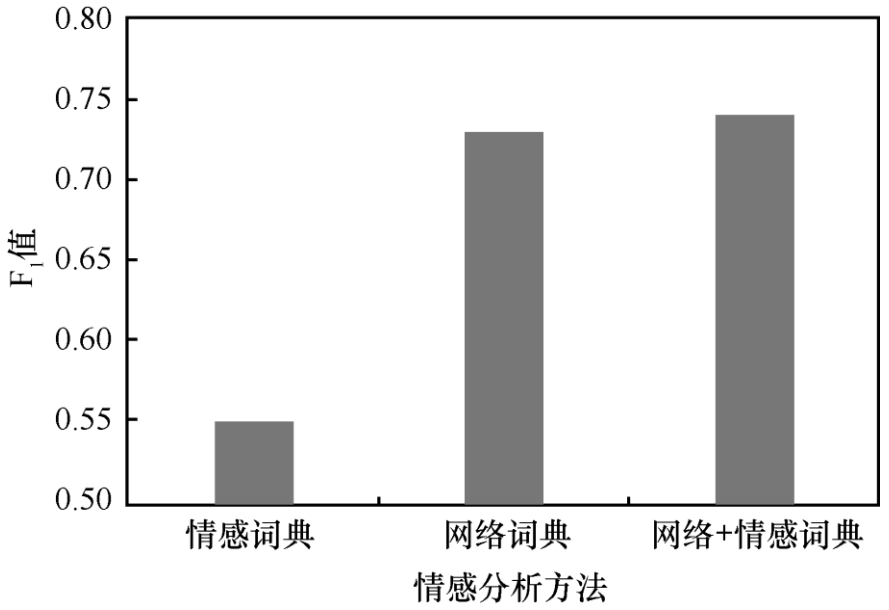


图4 基于词典方法F_1对比

使用基于词典的方法进行评论的情感分析，传统的情感词典效果并不是很好，F_1值在0.55左右，而采用一些来自评论自身的网络词汇构建的网络词典可达到较好的效果，F_1值能达到0.7以上。两者结合起来使用，比单独采用网络词典没有很好的提升效果。

4、评论情感分析

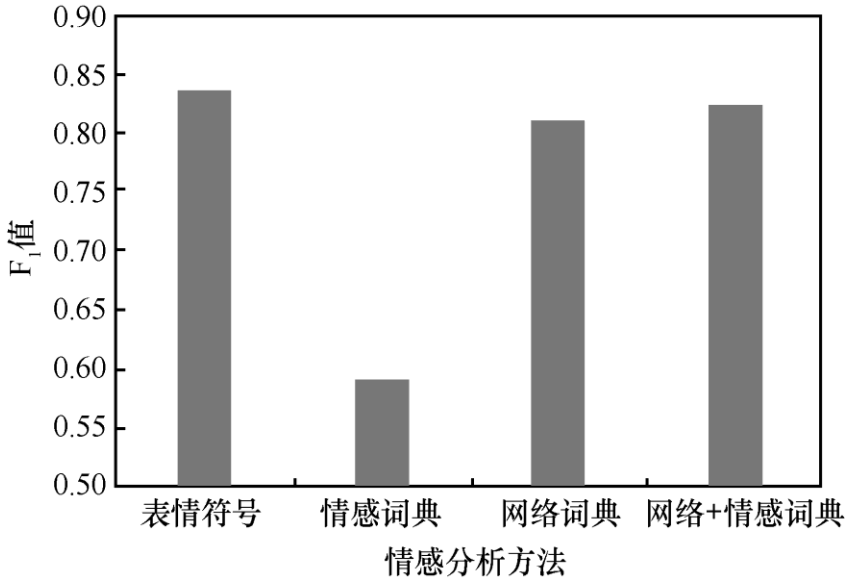


图4 基于表情符号方法情感分析F_1对比

对包含表情符号的评论采用基于表情符号的情感分析明显要好于采用传统的情感词典的情感分析。

因此，在评论的情感分析中，对包含表情符号的评论采用基于表情符号的情感分析加上对于不包含表情符号的评论采用基于网络词典的情感分析方法能够获得较好的结果。

目录

Contents

四、结论

1）可用活跃度分析微博用户活跃程度，紧密度和亲密度能够很好地发现用户间的关注程度以及潜在好友关系；

2）采用了特殊标点符号的微博话题提取方法，综合准确率达90.0%；

3）通过比较基于词典与表情符号和基于不同词典的评论情感分析，得出综合网络词典和表情符号的方法能取得更好效果。

Thank you