大数据与语言分析

华中师范大学　何婷婷

提纲

- 相关研究
- 博客语言专项调查
- 微博语言专项调查
- 年度十大网络用语

相关研究

- 互联网的出现与高度普及，拓展了语言生活的空间。在这个新的空间形式里，媒介的形式特征、网络用户的群体特征和特定网络应用的功能特征相互交织。
- 不少研究者试图以互联网空间中的“语言”为中心考察点来进行大规模的调查、分析。

网络语言相关研究

- Scott A. Golder 和 Michael W. Macy 2011年在 Science 杂志上发表了论文，通过对海量Twitter文本中的生活用语情况来调查在全球各地不同文化中的个体在每天和每个季节的情绪变化规律。通过对Twitter上生活用语的调查来表征用户的情绪表达，研究发现源于“睡眠、白昼长度、文化多样性”等特征引起的个体情绪变化变化存在周期恶化的规律。

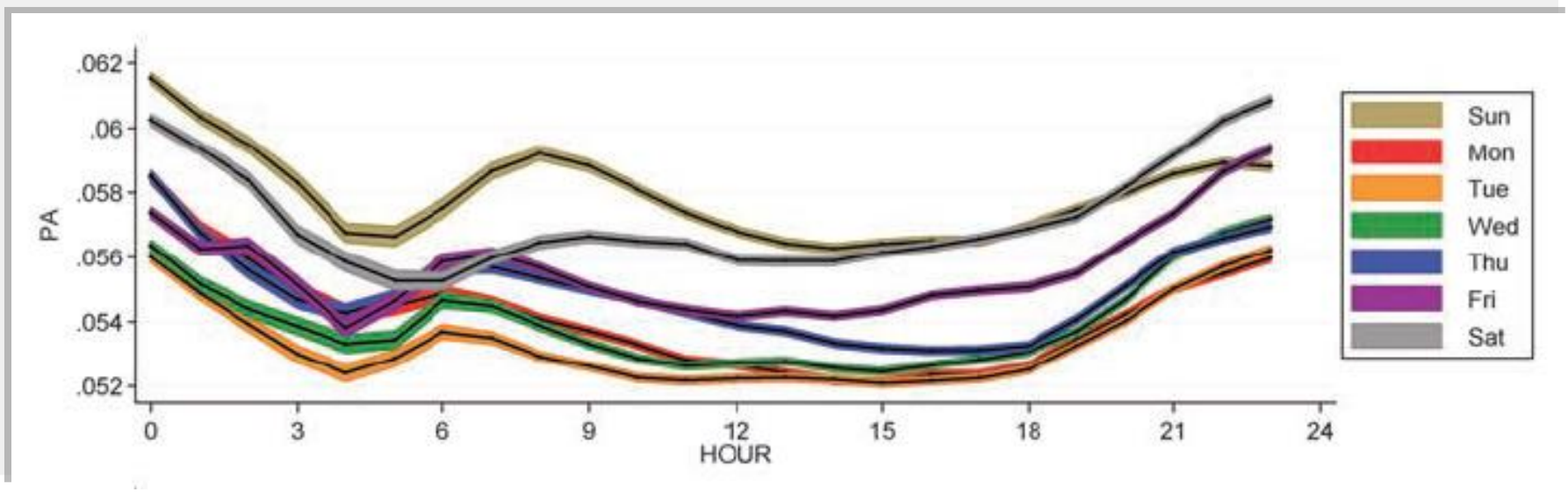
研究发现：

人们在早上醒来时心情较好，随着一天时间的推进心情逐渐恶化，这与睡眠的效果和昼夜节律一致，每个季节中人们的心情随着白天的时长的变化而变化。

人们在周末更快乐，但早上心情最好的时间会被延迟2 小时，这表明人们在周末会睡懒觉。

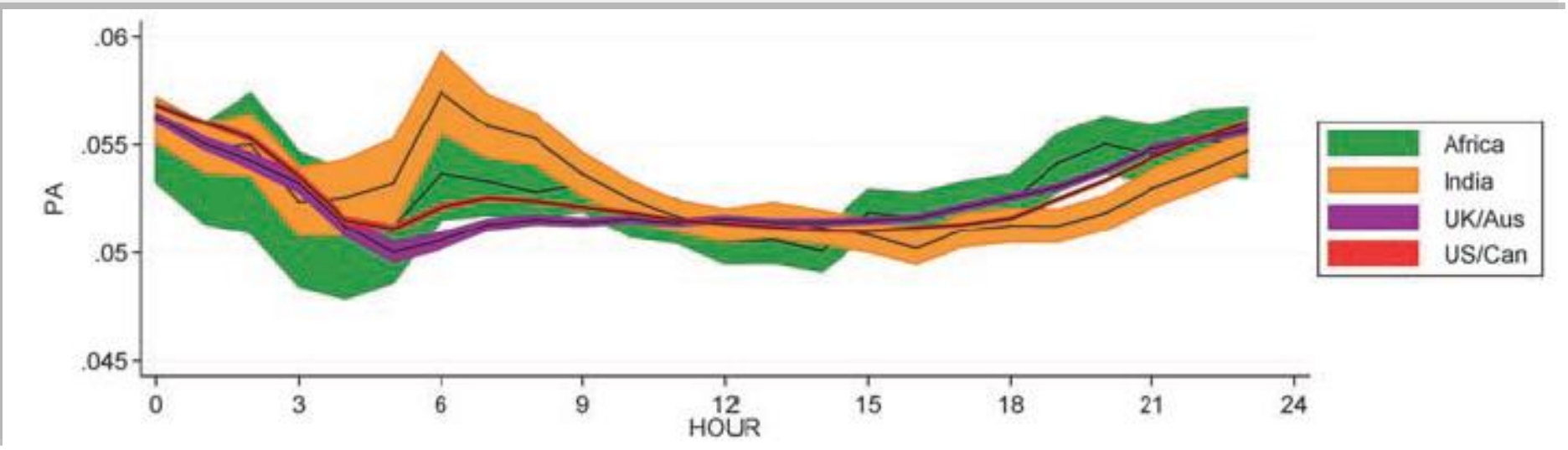
网络语言相关研究

- 人们在一个星期中每天的情绪变化情况



网络语言相关研究

- 不同地区的人们在每天的情绪变化情况



网络语言相关研究

- Sinan Aral & Dylan Walker 2012 年在 Science 杂志上发表了论文，将问题延伸到了网络营销的领域。

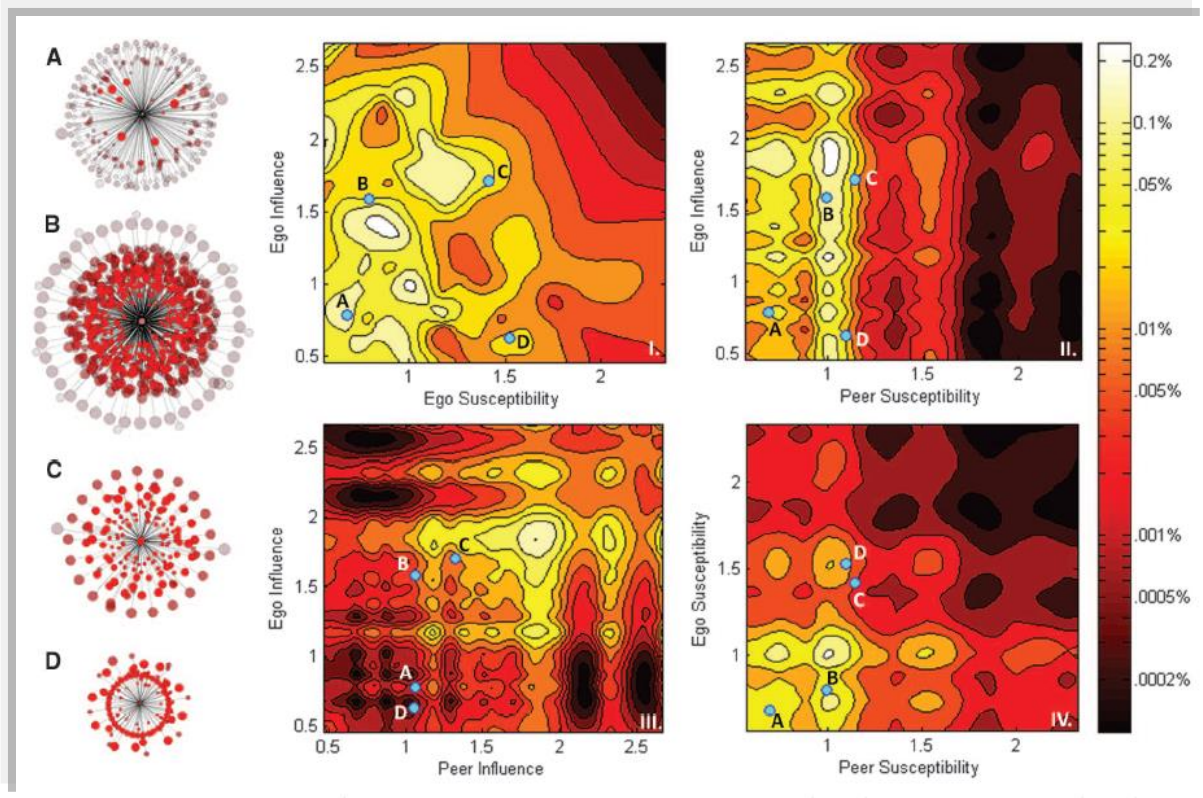
- 通过对130万Facebook用户生活用语的实验研究（对电影演员，导演的意见，对产品的评价），进行了用户生活用语与行为特征及性格特点、产品的推广相结合的多面探讨，研究发现年轻用户比年长用户更容易轻易受到其他人言论的影响。

实验表明：

- 年轻用户比老用户更容易受到影响；
- 男性比女性更有影响力，女性对男性的影响力大于她们对其它女性的影响力，已婚人士最不容易受到影响；
- 基于网络结构的影响力和传播分析表明，相对于没有影响力的人，有影响力的人通常不易受到影响，他们在网络上聚集，而易感个体通常没有这样做；
- 这表明拥有有影响力朋友的有影响力的人，有助于在产品在网络中的传播。(要好好利用在座的各位宣传我们的工作）

网络语言相关研究

- Facebook 用户影响力分析



网络语言相关研究

- David &Sandra（2005）通过关注青少年博客的语言特征来分析不同性别的青少年的用语特点，并以语言特征为聚焦点来探讨了不同性别青少年的在个性特点、情绪特征、性别认同感以及语义主题等指标值上的不同表现。
- Victor Savicki & Merle Kelley(2006)的研究可以看成是前一研究在更大规模上的扩充，通过对大规模网页中不同性别用语特征集的提取，来先验式判断网络个体的性别。

网络语言相关研究

- Andrew J &Yla R& James W（2010）实现了网络语言生活调查和社会心理学的有机结合。研究基于社会网络（SNS）带来的群体分层效应，分析群体中个体用语的特征在与日常生活语言比较的基础上探讨了同类型人群在社会关系、心理和情绪上的共同点.

2009年博客语言专项调查

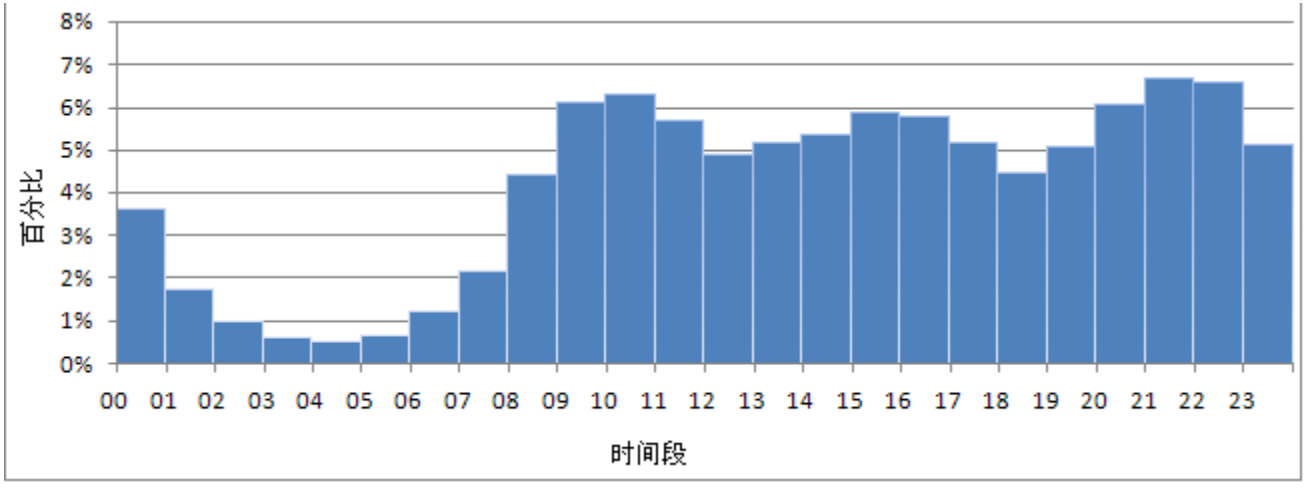
新浪博客　http://blog.sina.com.cn

搜狐博客　http://blog.sohu.com

- 17万个博客用户

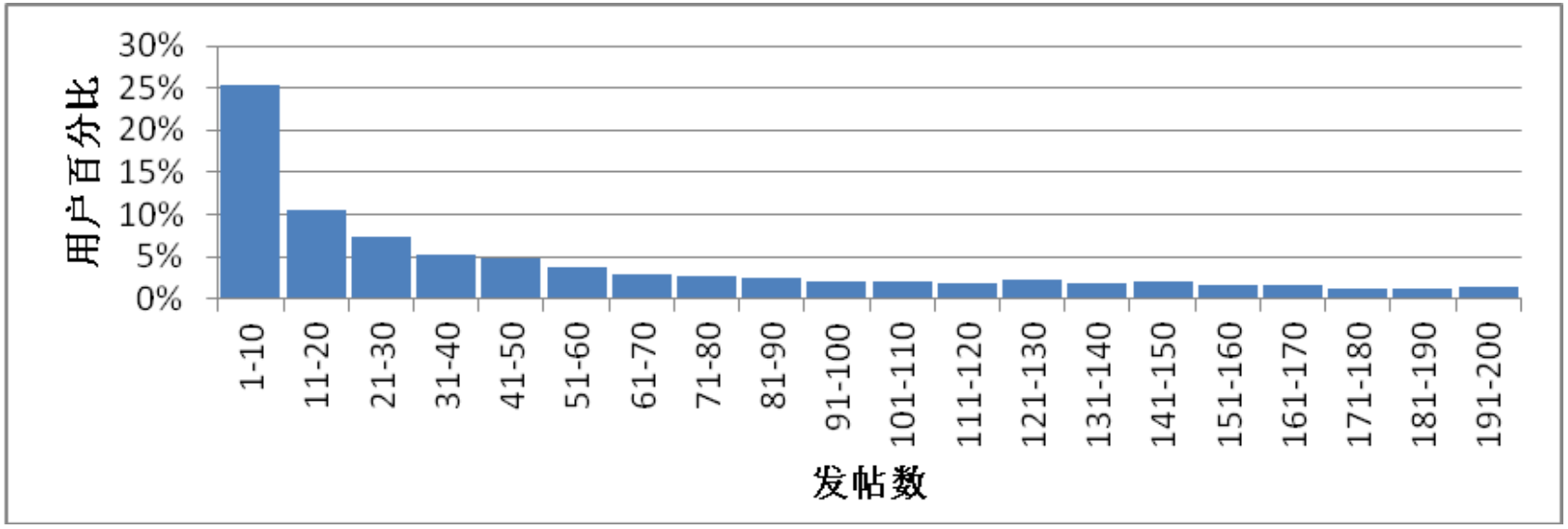
  全年发布的共计1千2百多万个博客帖，120多亿字符次，近百亿汉字。

- 调查内容包括博客用户发帖情况、博客用字用语情况以及博客标签使用情况。


8%
7%
6%
5%
4%
3%
2%
1%
0%
百分比
00 01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20 21 22 23
时间段

博客用户行为特点

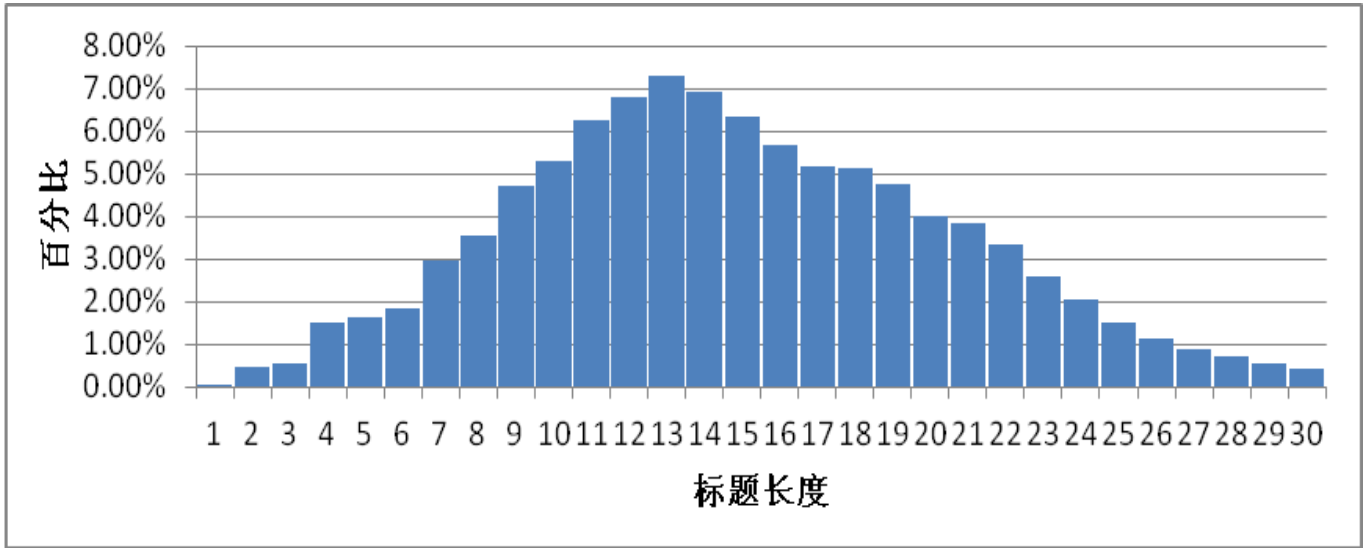
- 用户发帖量统计



各个发帖数量段的博客用户数分布（发帖数小于或等于**200**的用户）

博客用户行为特点

- 博客帖标题长度分布



长度小于等于30的博客帖标题长度分布

2009凤凰博客：博客作者性别语言分析

- 用字上的差异分析
  - 相对男性作者来说，女性作者使用较多的前20个字为:

    我，不，好，了,很,你,天,是,想,心,爱,的,么,都,真,那,得,还,小,去

  - 相对女性作者来说，男性作者使用比较多的前20个字为:

    国,中,用,大,年,球,作,之,学,本,文,行,这,者,成,而,方,业,主,于

博客作者性别语言状况

- 用词上的差异分析
  - 男女语文词使用对比

名词		动词		形容词	
男性	女性	男性	女性	男性	女性
问题	时候	为	想	高	好
社会	妈妈	使用	要	新	小
公司	人	进行	会	大	真
国家	朋友	出	喜欢	成功	开心
技术	女人	使	爱	强	幸福
球	心	打	说	不同	可爱
系统	爱情	可能	去	基本	快乐
游戏	心情	成为	觉得	重要	老
文件	生日	认为	看	伟大	亲爱
历史	爸爸	用	知道	低	久
兄弟	事情	提供	吃	一般	累
程序	日子	如	让	巨大	美丽
企业	男人	就是	哭	著名	难过
文化	事	进入	不	具体	漂亮
网络	家	需要	买	强大	坚强

博客作者性别语言状况

- 用词上的差异分析
  - 男女专有名词使用对比

人名		地名		机构名	
男性	女性	男性	女性	男性	女性
齐达内	安妮	中国	丽江	微软	麦当劳
毛泽东	许飞	美国	巴黎	中国队	意大利餐厅
孔子	金牛	日本	西单	联合国	韩国料理店
黄健翔	王菲	意大利	云南	意大利队	华师大
鲁迅	刘诲	德国	桂林	新浪	华山医院
舒马赫	陈怡川	北京	九寨沟	德国队	韩国餐厅
亨利	张爱玲	台湾	松江	中国政府	湖南台
姚明	尚雯婕	上海	泰国	法国队	上海美术电影制片厂
曹操	张小娴	欧洲	乌镇	国务院	信乐团
刘备	厉娜	阿根廷	阳朔	网易	电子科技大学
布什	思远	法国	海南	雅虎	港式茶餐厅
李敖	晓波	朝鲜	厦门	北京大学	悉尼歌剧院
巴乔	菲菲	巴西	淮海路	清华大学	上海卫视
格罗索	李俊基	米兰	三亚	教育部	火锅店
马克思	柯南	英国	大理	美国政府	朝阳公园
乔丹	孔吉	香港	新疆	澳大利亚队	泰国餐厅
悟空	俞思远	荷兰	外滩	中国共产党	人大附中
周恩来	金三顺	亚洲	王府井	美国队	世纪公园

博客作者性别语言状况

- 字、词多样性上的差异分析

男女用字覆盖率分布

覆盖率	80%	90%	99%	100%
男性字种数	546	1 005	2 964	20 916
女性字种数	495	953	2 912	13 558

男女用词覆盖率分布

覆盖率	50%	80%	90%	99%
男性词种数	279	3 578	10 479	76 087
女性词种数	201	2 562	7 660	55 699

数据表明，男性作者在汉字、词语的使用上更加具有多样性。

2010年博客语言调查

- 2010年调查语料来自新浪、网易和搜狐三家网站，这些网站都公布了名博列表。本次调查统计了这些列表中的1 929个博客用户全年发布的共计176 089个博客帖。调查内容包括博客用户发帖情况、博客用字用语情况以及博客标签使用等。

2012年度博客语言调查

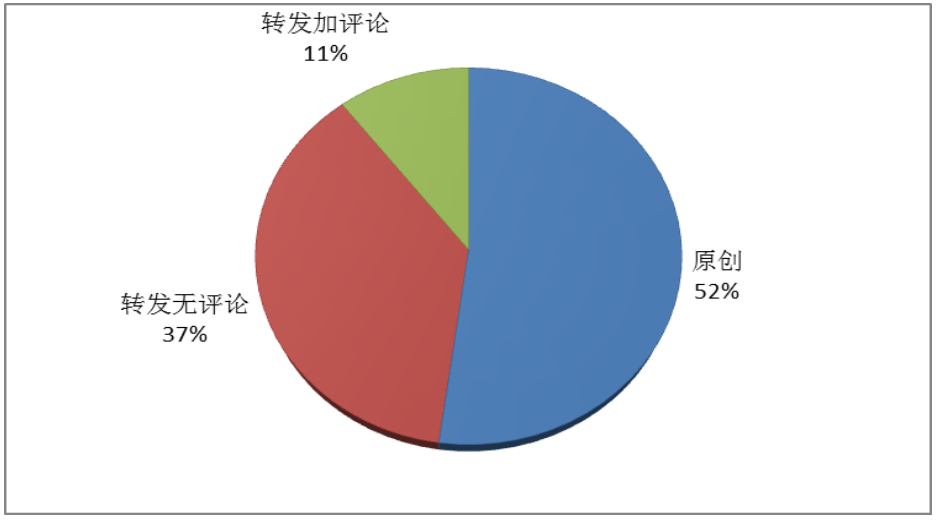
- 调查语料来自新浪、网易和搜狐三家网站，包括10 875个博客用户全年发布的共计240 763个博客帖。
- 在这些博客用户中，有3 281个博客用户是由各个网站推荐的知名博客，他们共发布了117 497个博客帖；另有7 594 个普通博客用户，他们共发布了123 266个博客帖。

- 知名博主的博客帖中关心的话题更多地与经济、政治等相关；而普通博主的博客帖中更多地与生活、情感等个体行为相关。

- 人名、地名、机构名、事件名
- 基于标签的热点事件、人物等

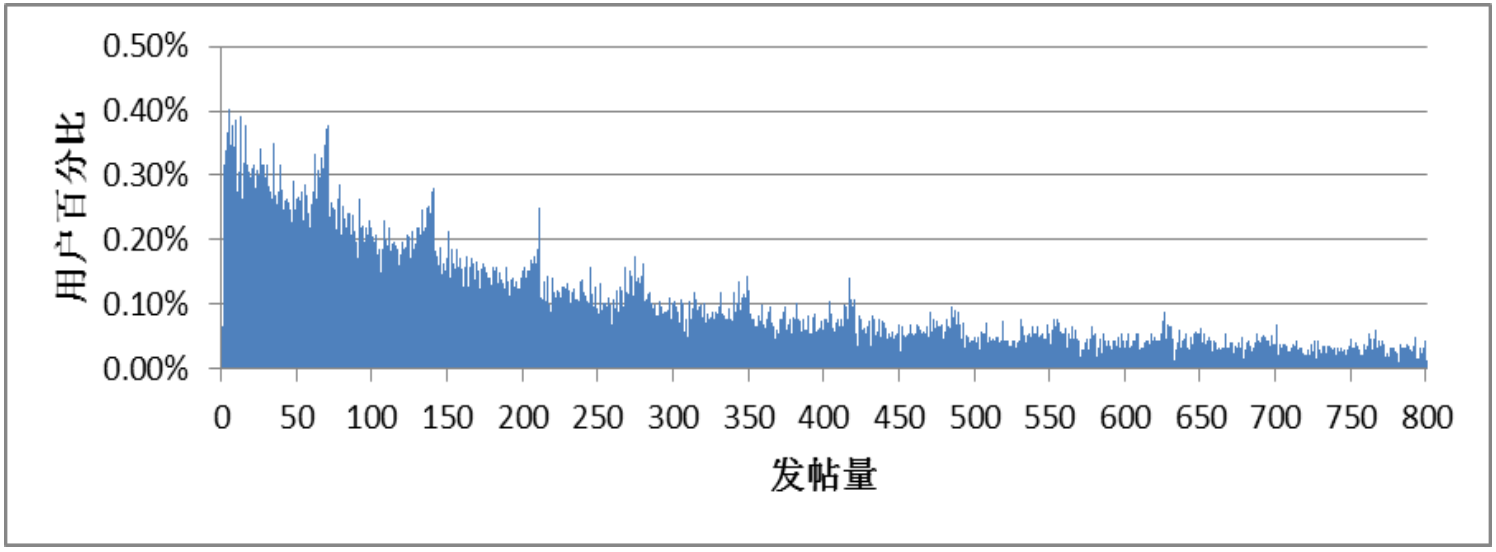
2013微博语言专项调查

2013年，本次调查的45,000个微博用户全年共发布20,307,537个微博帖。我们分别统计了原创帖、转发无评论帖和转发加评论帖的总量、比例。



用户发帖行为分析

- 用户发帖量分析



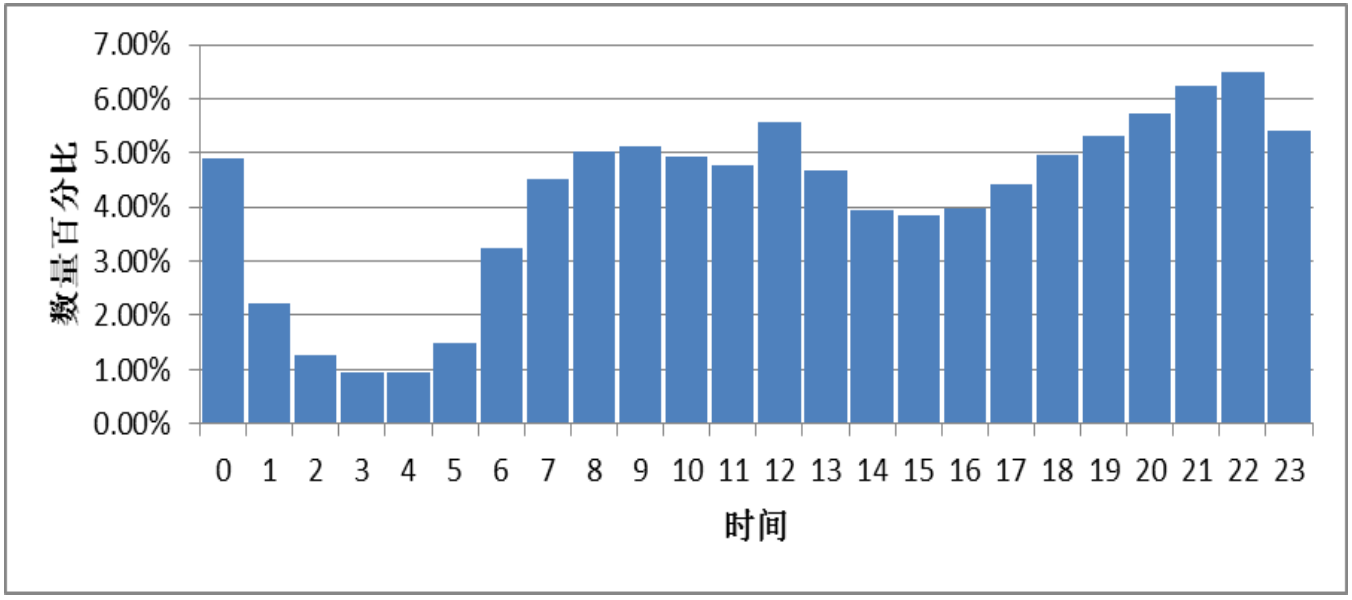
各个发帖量数量段的微博用户数分布图(发帖量0-800之间)

- 占总数约88%的微博用户，发帖总量只占总数的54%，而剩下的约46%微博帖却由12%的用户所发。
- 在调查的微博用户中，发帖量最多的为 24853篇，发帖量中位数为238篇。
- 发帖最多的前10个用户中，有9个为普通个人用户，这些用户的知名度并不高，我们通常把这类用户称为“微博控”；剩下的1个为机构微博用户，主要的目的为微博营销。

- 为了更好的了解不同微博群体的发帖行为，我们根据发帖量把用户分为3个不同的群体：
- (1) 活跃用户，即发帖量最多的前9000个用户(占用户总数的20%)；
- (2) 不活跃用户，即发帖量最少的9000个用户；
  (3) 普通用户，为其余的27000个用户。
- 统计数据表明，活跃用户平均每天发布3.8个微博帖，普通用户平均每天发布0.8个微博帖，不活跃用户平均每月发布2.7个微博帖。

用户发帖行为分析

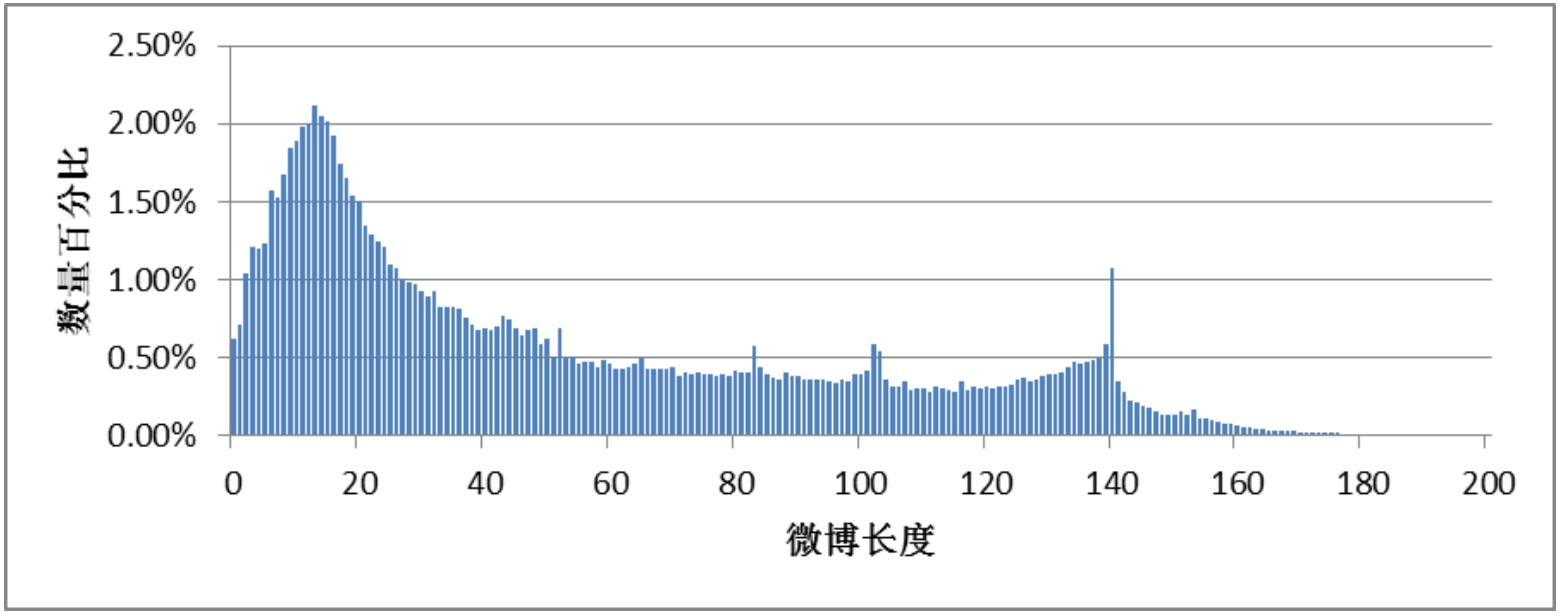
- 用户发帖时间段统计



发帖时间分布图

用户发帖行为分析

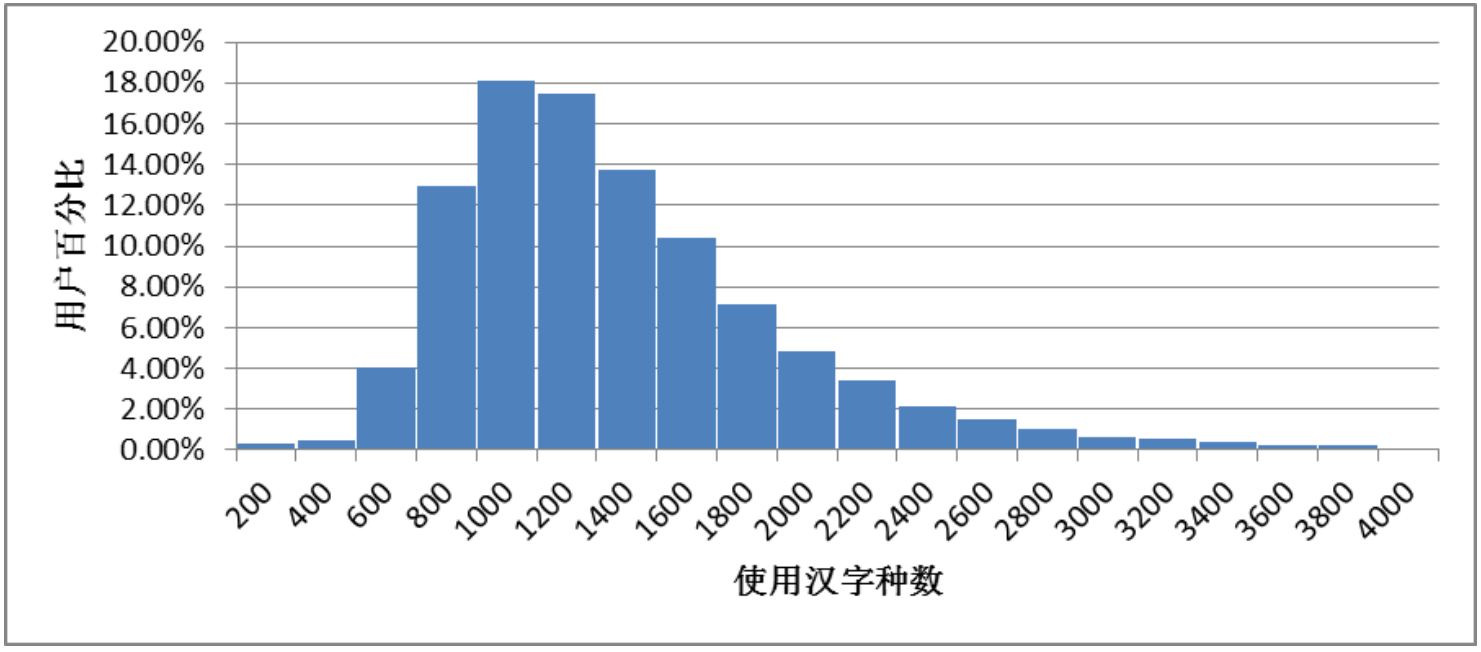
- 微博帖长度分布



原创微博帖长度分布图(长度为1-200之间)

- 对于原创帖而言，长度小于等于50个字符的约占总数的60%，长度小于等于100个字符的约占总数的80%，长度大于100个字符的仅占总数的20%。

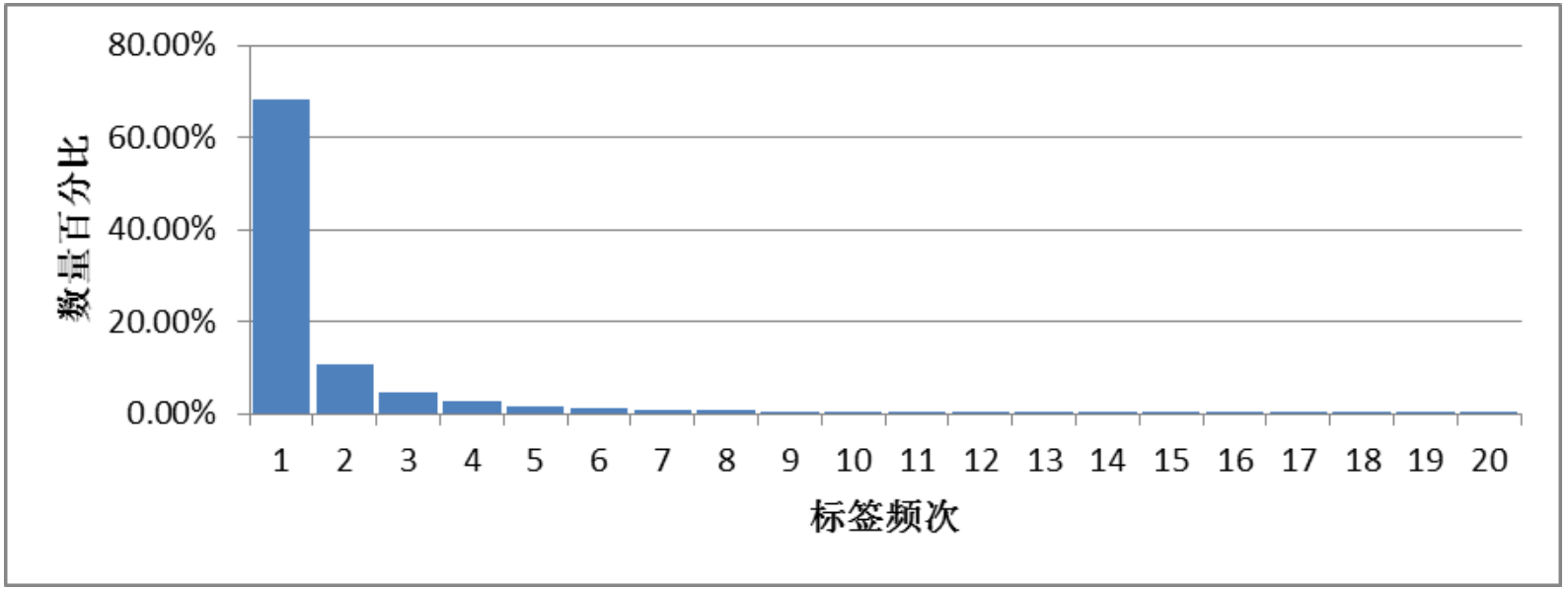
微博用户汉字使用情况



使用汉字种数分布图 (字种数在0-4000以内)

- 约36% 用户使用的汉字字种数在1000以内
- 约98% 用户使用的汉字字种数在3000以内
- 大部分的微博用户倾向于使用简单常见的汉字，从侧面反映出微博语言口语化和日常化的特点。（“好看 ” “漂亮”“美丽”）

话题标签情况



标签使用频次分布图 (频次在0-20以内)

2012、2013十大网络用语发布

谢谢！