



# 微博用户特征与行为的大数据挖掘

张华平 博士 副教授

@ICTCLAS张华平博士



<http://www.nlpir.org/>

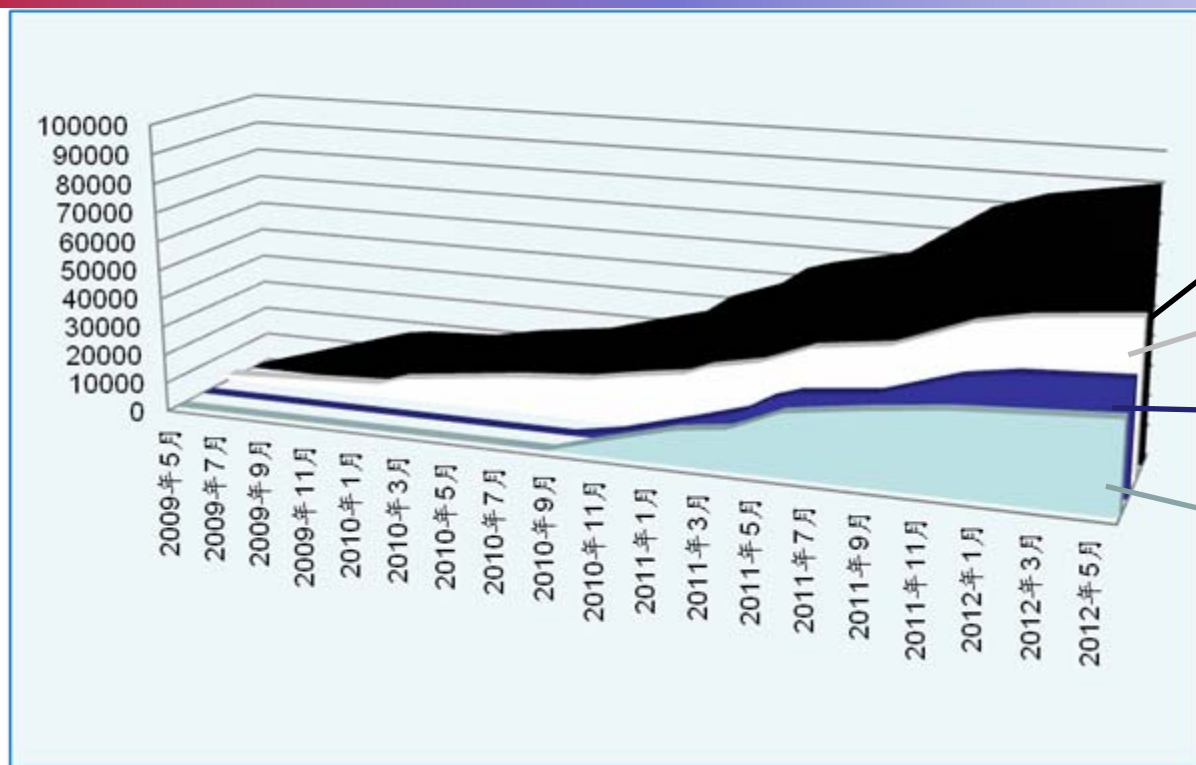
第九届中国网络科学论坛

2013/4/25





# 社交网络应用的迅猛发展



Facebook: 9亿

Twitter: 5亿

腾讯微博: 3.3亿

新浪微博: 3亿

Facebook上线不足8年，已拥有超过**9亿**的用户，是第三大“人口国”

**CNNMoney**  
A Service of CNN, Fortune & Money

FORTUNE

Money

## Facebook surpassed 900 million users

By David Goldman @CNNMoneyTech April 23, 2012: 4:31 PM ET

NEW YORK (CNNMoney) -- Facebook surpassed 900 million active users last month, according to a regulatory filing, helping the social network post more than \$1 billion in sales in the first quarter.



BEIJING INSTITUTE OF TECHNOLOGY



# 大数据 vs. 小数据

小  
小  
小小小小小小小小小小  
小  
小  
小小小小小小小小小小  
小小小小小小小小小小小小



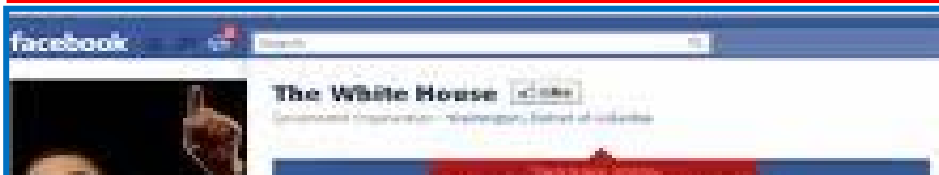
# 纲要



# 在线社交网络对生活方式的影响



**竞选：**奥巴马通过社交网络进行助选、民意调查



**谣言：**我国民众受微博谣言蛊惑而抢盐，影响社会稳定





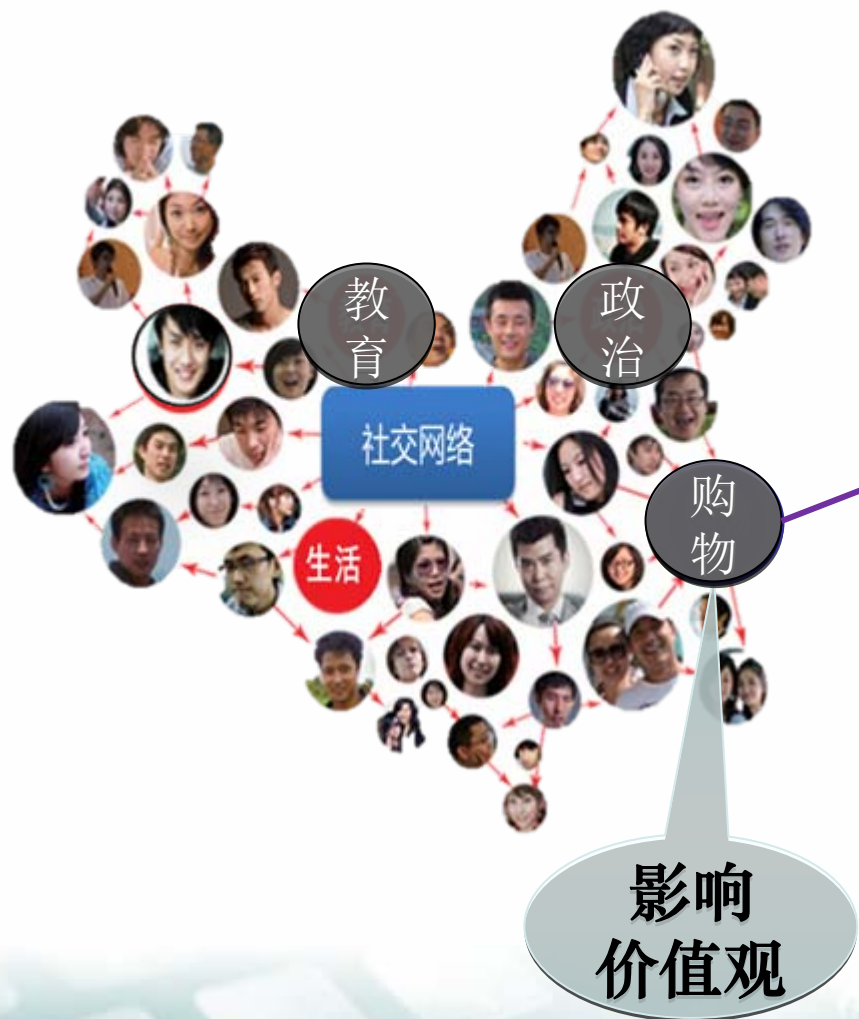
# 在线社交网络对生活方式的影响



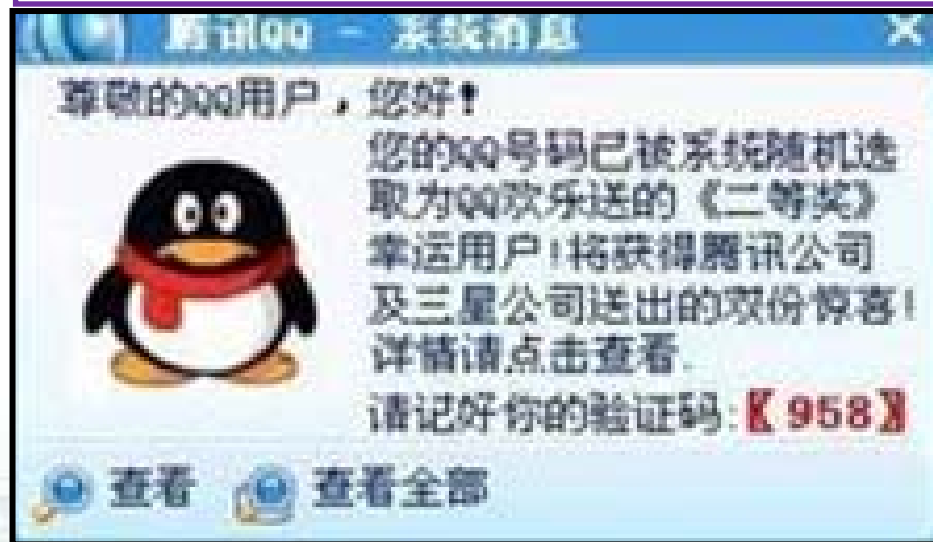
**暴力渲染：**邪恶势力通过社交网络教育青年人崇尚暴力，鼓吹破坏性的个人英雄主义



# 在线社交网络对生活方式的影响



**欺诈：**不法分子通过聊天工具发布虚假信息、利用在线购物平台欺诈顾客。



# 在线社交网络对生活方式的影响



**色情诱惑：**低俗团伙利用社交视频、即时通信等渠道传播色情



影响  
人生观



# 在线社交网络为突发事件推波助澜

影响

利比亚Facebook革命



购物

Twitter煽动 “茉莉花革命”



影响  
价值



北京理工大学  
BEIJING INSTITUTE OF TECHNOLOGY



# 社交网络与信息安全解读

➤ 宏观安全：国家基础数据的战略安全、宏观决策安全；

➤ 微观安全：个人隐私、个性与行为模式的安全；

宏观特  
征数据  
挖掘

微观个  
性与行  
为建模

话题与  
情感内  
容分析

➤ 内容安全：信息内容的安全、意识形态安全。



北京理工大学  
BEIJING INSTITUTE OF TECHNOLOGY



# 宏观特征大数据挖掘说明

- 抓取技术：模拟浏览器；持续两年，数据存在一定滞后性，但不影响宏观规律
- 抓取策略：给定一批种子，只抓取其关注对象，确保微博用户数据的质量；
- 字段包括：性别/地址/粉丝数/关注数/教育信息/工作经历/生词/简述
- 清洗后的数据规模为1700万(摒除大量机器自动生成的僵尸用户及休眠用户)。样本=总体
- 部分数据进行隐私处理后发布在 [www.nlpir.org](http://www.nlpir.org) 上。





# 微博用户数据样本

id	url	name	sex	birthday	address	fansNum	summary	wbNum	gzNum	blog	realName
10315	http://weibo.com	老军鹏	男	<NULL>	北京 海淀区	209	他还没填写个人介	446	157	<NULL>	<NULL>
10318	http://weibo.com	王海荣	男	<NULL>	<NULL>	600	<NULL>	<NULL>	<NULL>	<NULL>	<NULL>
10362	http://weibo.com	kanglicoco	女	<NULL>	<NULL>	454	懂点事儿的小清	837	57	<NULL>	<NULL>
10413	http://weibo.com	赵和	男	<NULL>	北京 海淀区	463	崇辱不惊，闲看庭	369	370	http://blog.sina.com	<NULL>
10469	http://weibo.com	张淼atSina	女	<NULL>	<NULL>	230	幸福的每一天	755	208	http://blog.sina.com	<NULL>
10514	http://weibo.com	闸北陆小洪	男	1985年10月9日	北京 海淀区	538	互撸娃，努力学习	2016	136	<NULL>	<NULL>
11022	http://weibo.com	吴军	男	1981年1月1日	广东 广州	938	他还没填写个人介	1174	432	http://blog.sina.com	<NULL>
11051	http://weibo.com	protobuf	男	<NULL>	北京 海淀区	2022	Protocol Buffers fans	0	0	<NULL>	<NULL>
11075	http://weibo.com	朱磊	男	<NULL>	北京 海淀区	575	微博招JAVA研发人	324	580	http://blog.sina.com	<NULL>
31790	http://weibo.com	杯中威士忌	男	<NULL>	黑龙江 哈尔滨	185	不求数量，只求质	524	89	<NULL>	<NULL>
32146	http://weibo.com	冰鱼孙靖儿	女	<NULL>	<NULL>	22	喜欢唱歌。喜欢写	4	3	http://blog.sina.com	<NULL>
32884	http://weibo.com	一抹湖水	男	<NULL>	山东 青岛	17	有人说高山上的湖	11	14	<NULL>	<NULL>
35277	http://weibo.com	晓鑫_A	男	<NULL>	<NULL>	447	<NULL>	<NULL>	<NULL>	<NULL>	<NULL>
35882	http://weibo.com	翰墨飘香66	男	1955年5月19日	北京 海淀区	131	真实感人	73	1593	http://blog.sina.com	<NULL>
40783	http://weibo.com	邹建_民生证券	男	1972年10月6日	四川 成都	356	爱业、敬业、专业	1327	795	http://blog.sina.com	<NULL>
41499	http://weibo.com	何汉三	男	<NULL>	<NULL>	8314	<NULL>	<NULL>	<NULL>	<NULL>	<NULL>
76577	http://weibo.com	XIE_LIN	男	天蝎座	北京	80	与你分享.....	456	30	<NULL>	<NULL>
79608	http://weibo.com	黑膠情	男	1971年11月11日	海外 意大利	617	世界不会在意你的	3469	211	http://blog.sina.com	<NULL>
79660	http://weibo.com	晴娃娃79660	女	1983年5月29日	<NULL>	1106	大量5、6位微博数	1319	1445	http://weibo.com/79	<NULL>
82506	http://weibo.com	人大附中	女	<NULL>	<NULL>	58	她还没填写个人介	0	4	<NULL>	<NULL>
95095	http://weibo.com	耿一正	男	9月3日	北京 朝阳区	33477	以前不等于现在、	357	149	http://blog.sina.com	耿一正
98122	http://weibo.com	团购网址导航网	男	魔羯座	广东 深圳	1115	创炜基团购网址ht	794	532	http://blog.sina.com	<NULL>
99001	http://weibo.com	汪洋洋	男	<NULL>	北京	510	命里有时终须有，	655	281	<NULL>	<NULL>
101713	http://weibo.com	京涛Hi浪	男	<NULL>	北京 海淀区	488	脑子进水了	2712	325	<NULL>	<NULL>
103500	http://weibo.com	张宴	男	1985年5月19日	北京 海淀区	84283	专注于架构设计、	309	1896	http://blog.s135.com	张宴
103558	http://weibo.com	李雁春	女	<NULL>	<NULL>	999	难道就此开始写东	950	238	<NULL>	<NULL>
103759	http://weibo.com	荀志锋	男	<NULL>	北京 海淀区	796	择高处立，就平处	2374	1550	<NULL>	<NULL>
103778	http://weibo.com	高勇	男	<NULL>	北京 海淀区	464	海阔凭鱼跃，天高	1537	346	<NULL>	<NULL>
104104	http://weibo.com	夏思	男	<NULL>	北京 海淀区	512	忠实的#电影#美剧	2601	299	http://blog.sina.com	<NULL>
104508	http://weibo.com	乾中	男	1982年2月12日	北京 海淀区	479	得闲饮茶	1061	196	http://blog.sina.com	邓乾中
104541	http://weibo.com	此微薄不用了	男	<NULL>	北京 海淀区	91	更新尽在：http://t	18	84	http://blog.sina.com	<NULL>

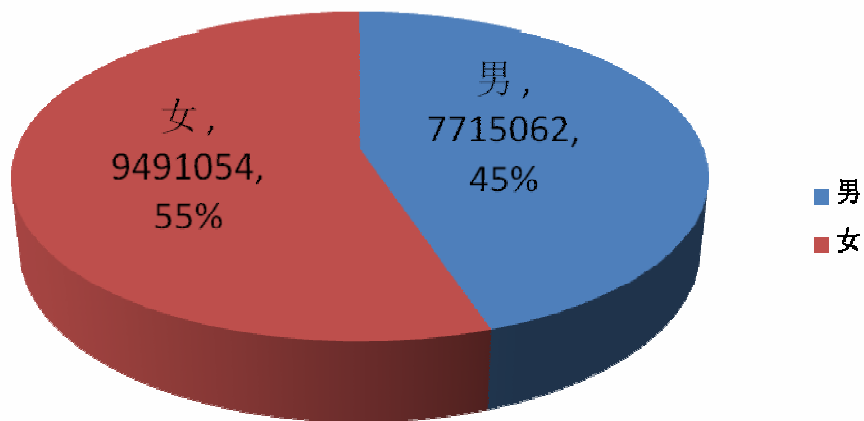




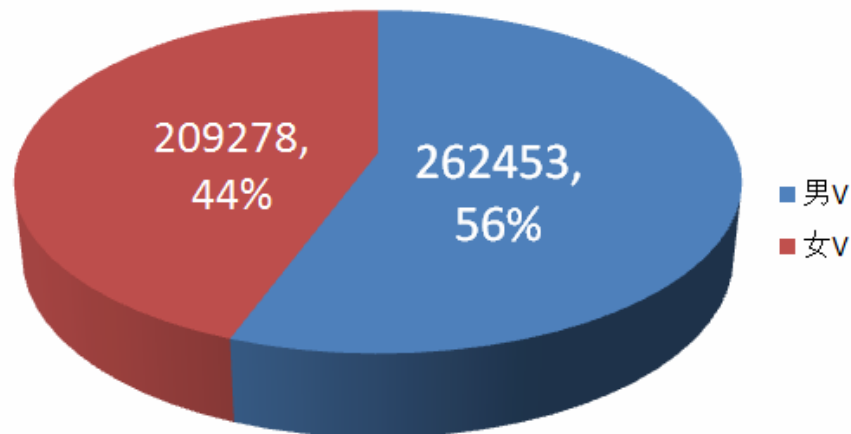
# 性别比例分布

性别	人数
男	7715062
女	9491054
合计	17206116

男女比例图表



认证用户男女比例

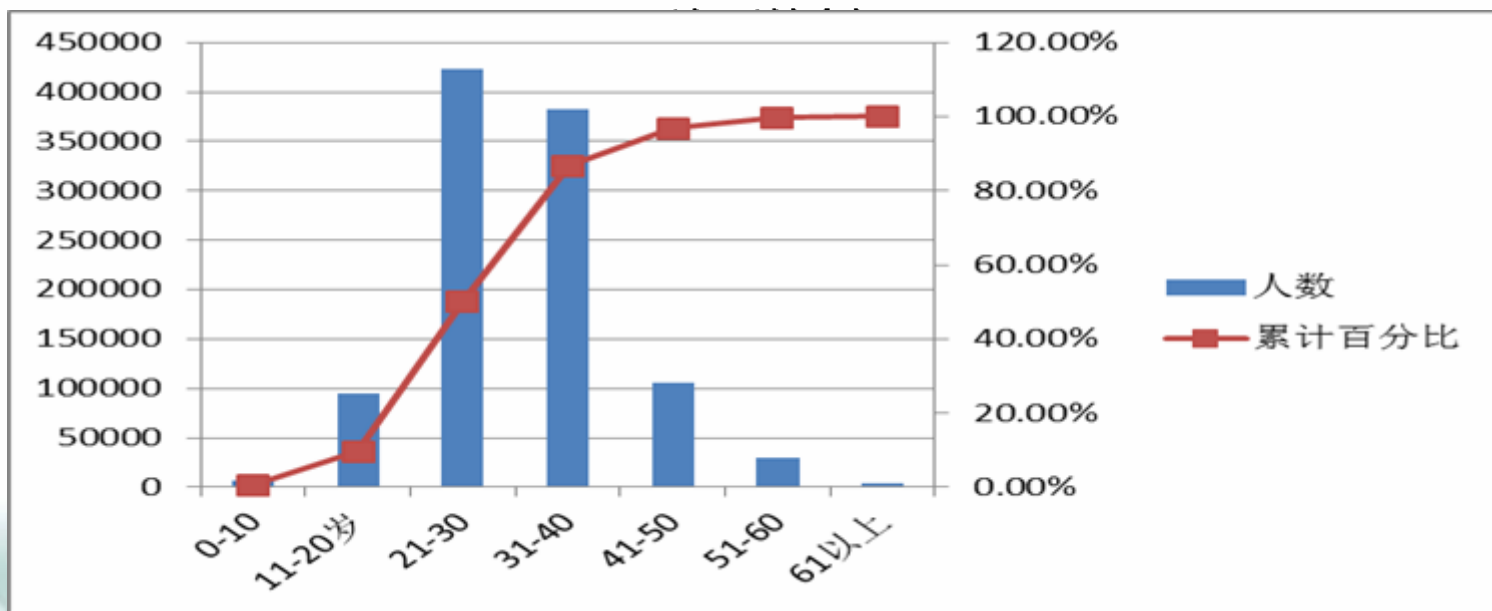






# 不同类型用户的分布

认证级别	人数	备注
无认证	16746493	97.26%
1级认证	201423	认证个人
2级认证	23150	政府机构
3级认证	246228	企业等机构
4级认证	931	焦点人物

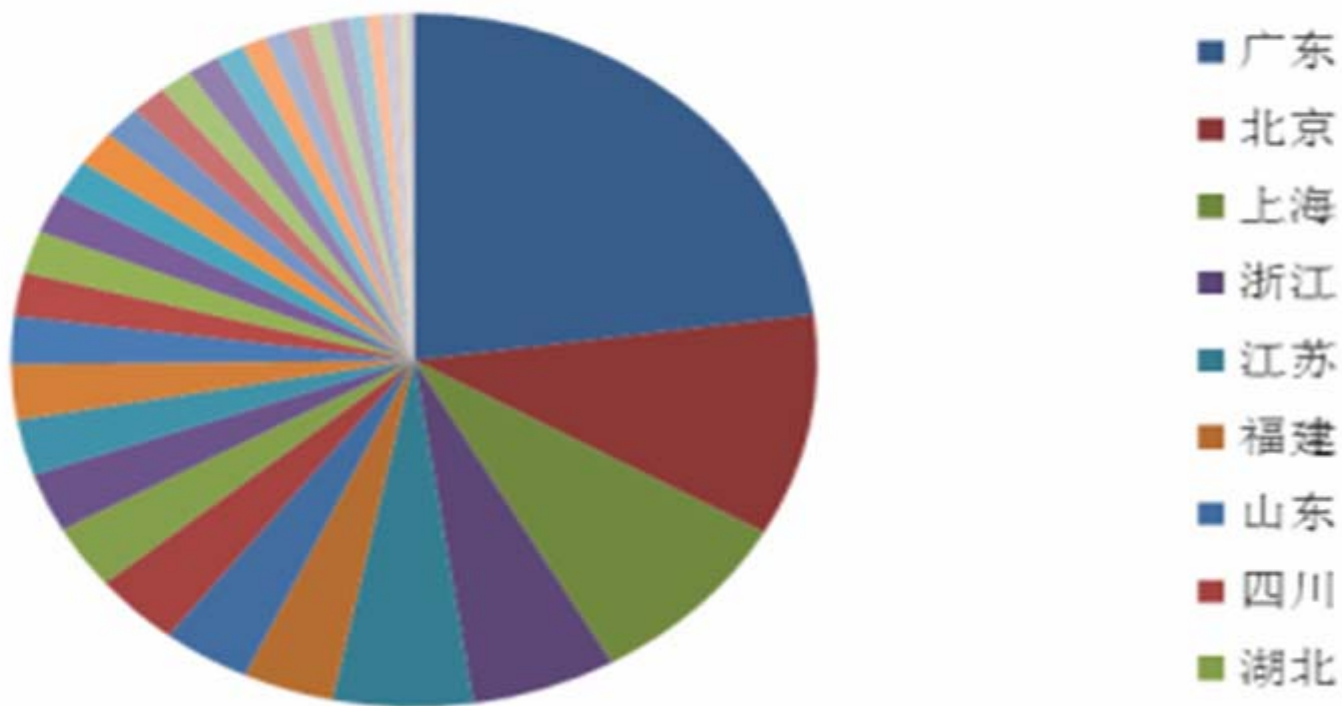




# 不同地区的微博用户总数

总用户数

Geometric Distribution



江  
地区

吉

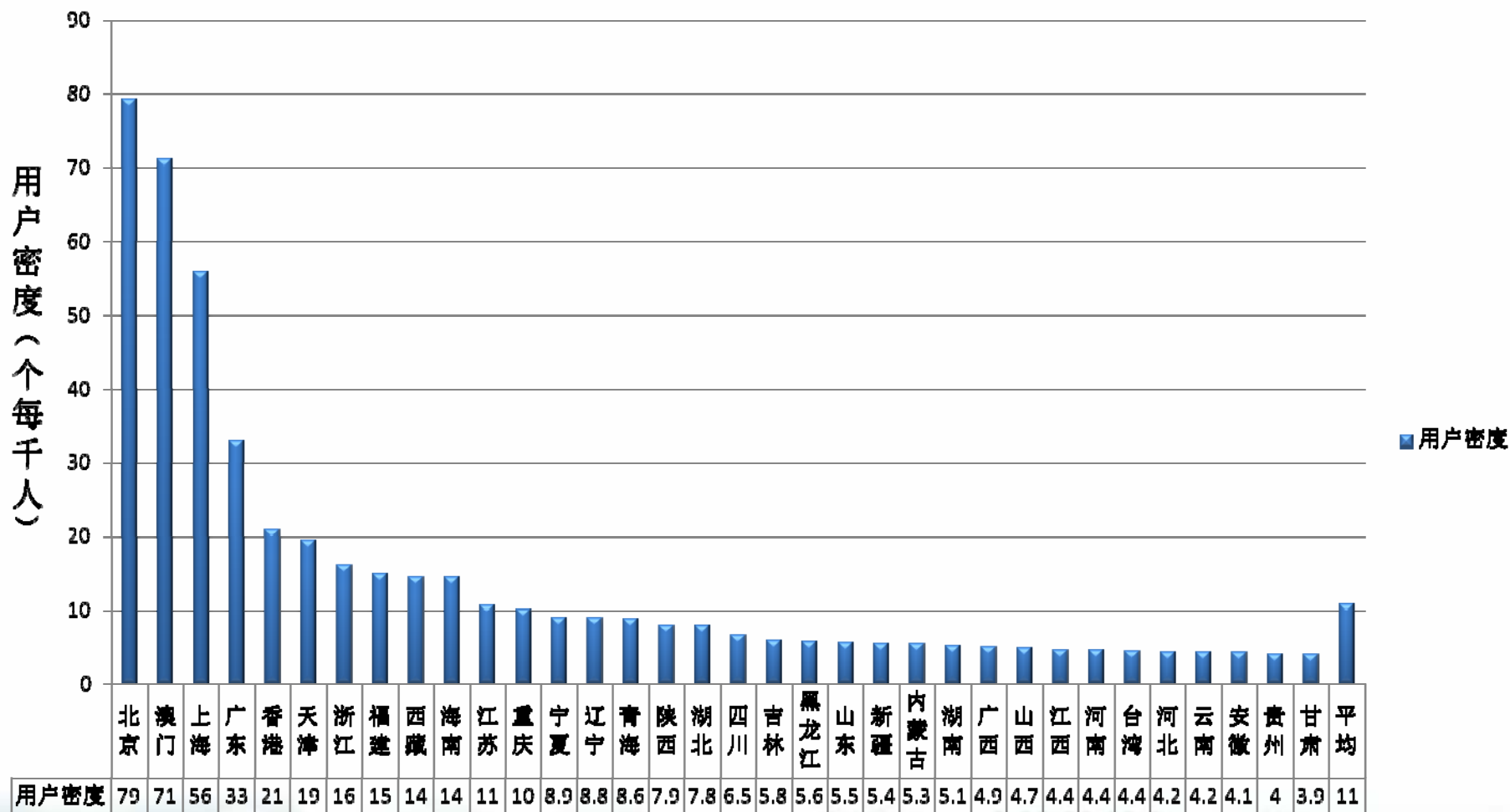


北京理工大学  
BEIJING INSTITUTE OF TECHNOLOGY



# 不同地区的微博用户密度

## 用户密度



地区

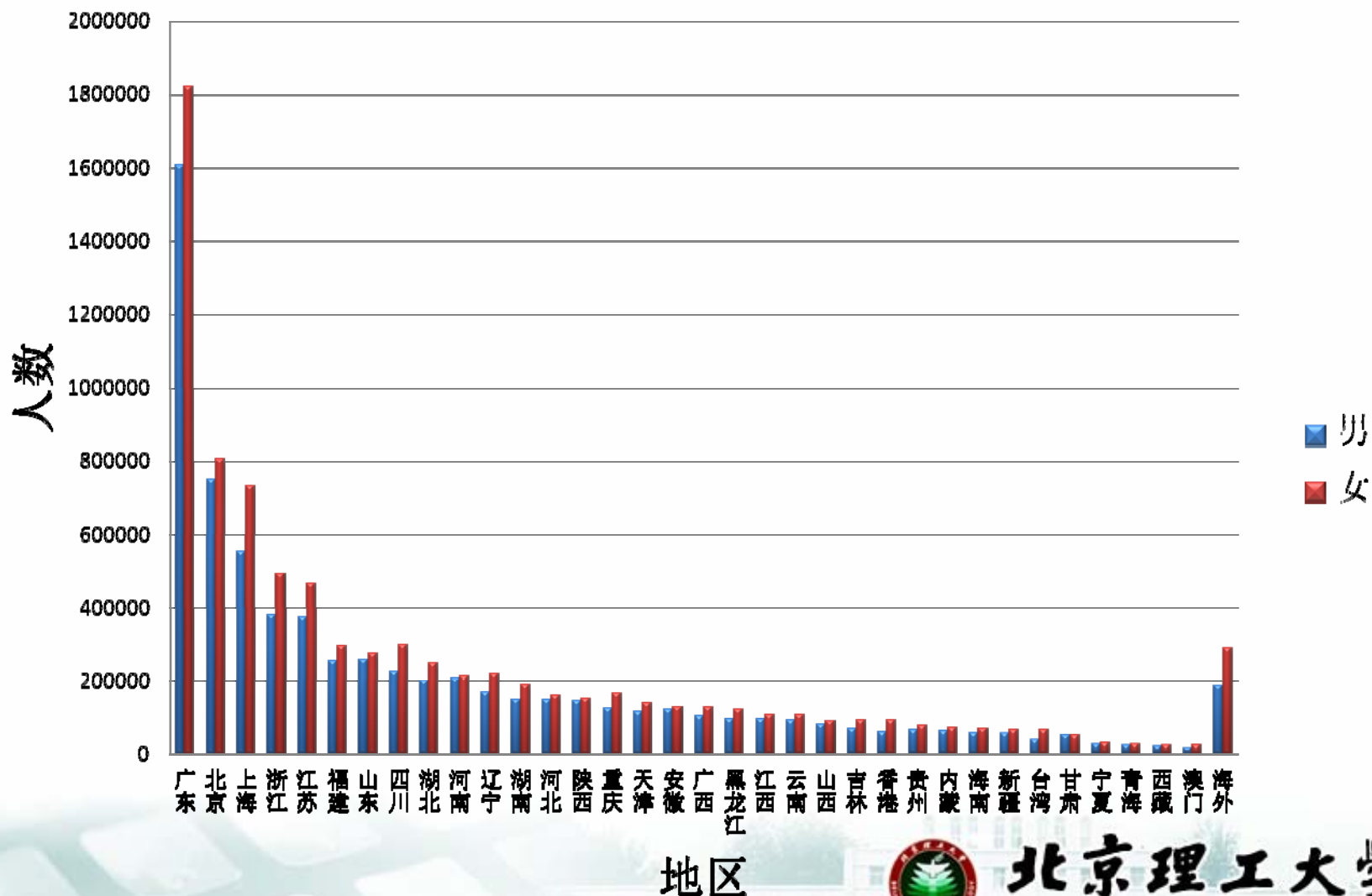


北京理工大学  
BEIJING INSTITUTE OF TECHNOLOGY



# 性别/区域比例联合分布

## 地区——性别比例联合分布



北京理工大学  
BEIJING INSTITUTE OF TECHNOLOGY



# 省市区划内微博用户数与GDP正相关

北京区划	绝对人数	占总体比列
朝阳	525527	39.91%
海淀	353901	26.88%
东城	108112	8.21%
西城	105200	7.99%
顺义	66417	5.04%
丰台	36845	2.80%
昌平	25694	1.95%
通州	24724	1.88%
石景山	21816	1.66%
大兴	21816	1.66%
房山	7756	0.59%
密云	7756	0.59%
平谷	6302	0.48%
怀柔	2424	0.18%
门头沟	1454	0.11%
延庆	969	0.07%
总计	1316713	100.00%

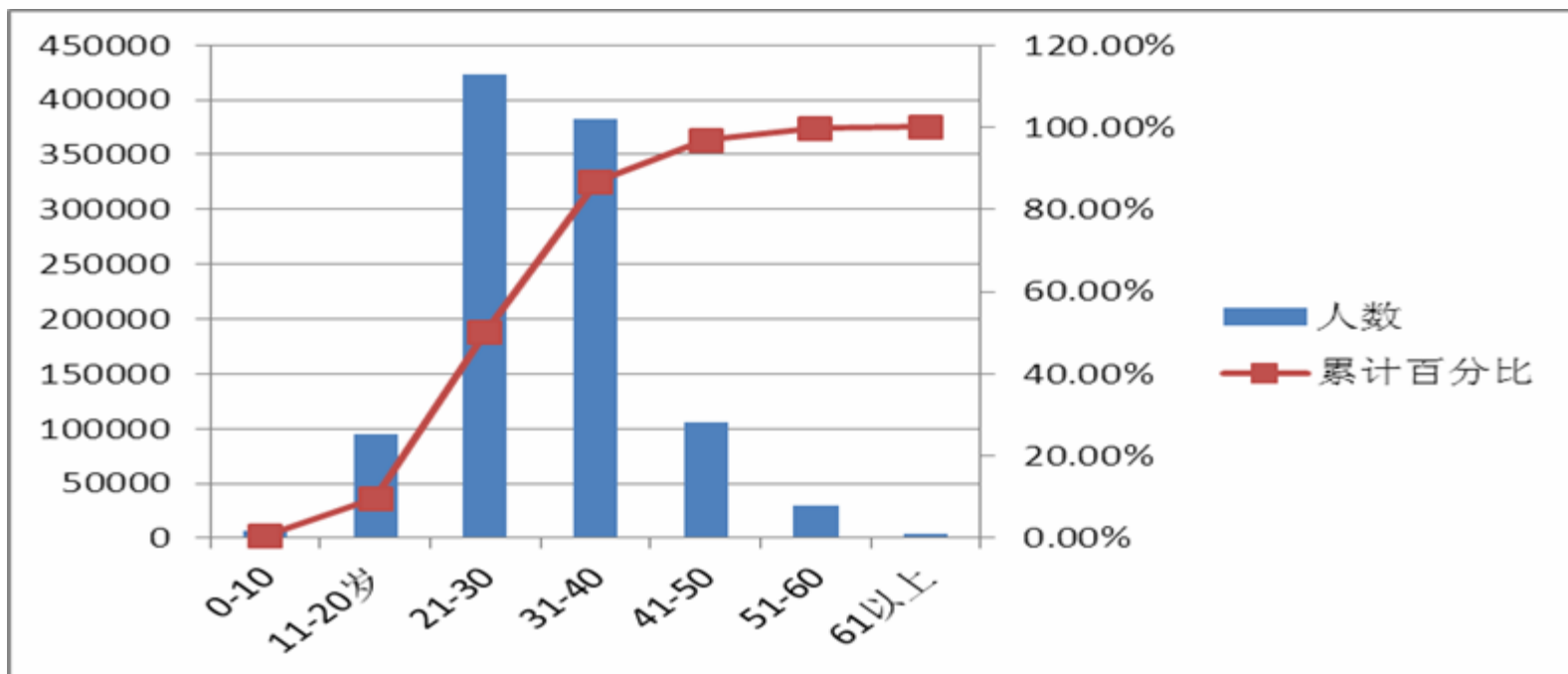
区 县	地区生产总值		
	2010	2009	增长速度(%)
全 市	14113.6	12153.0	10.3
朝 阳 区	2804.2	1122.4	9.0
海 淀 区	2771.6	1815.6	13.3
西 城 区	2057.7	2380.4	17.8
东 城 区	1223.6	627.4	17.1
顺 义 区	867.9	248.7	18.8
丰 台 区	734.8	2446.9	13.3
昌 平 区	399.9	293.5	26.6
房 山 区	371.5	278.9	23.6
通 州 区	344.8	690.2	25.7
大 兴 区	311.9	342.4	16.8
石景山区	295.5	271.2	15.0
怀 柔 区	148.0	74.8	15.6
密 云 县	141.5	131.4	12.6
平 谷 区	117.9	107.0	10.2
门头沟区	86.4	119.5	18.3
延 庆 县	67.7	61.5	10.1
北京经济技术开发区	698.6	592.5	17.9





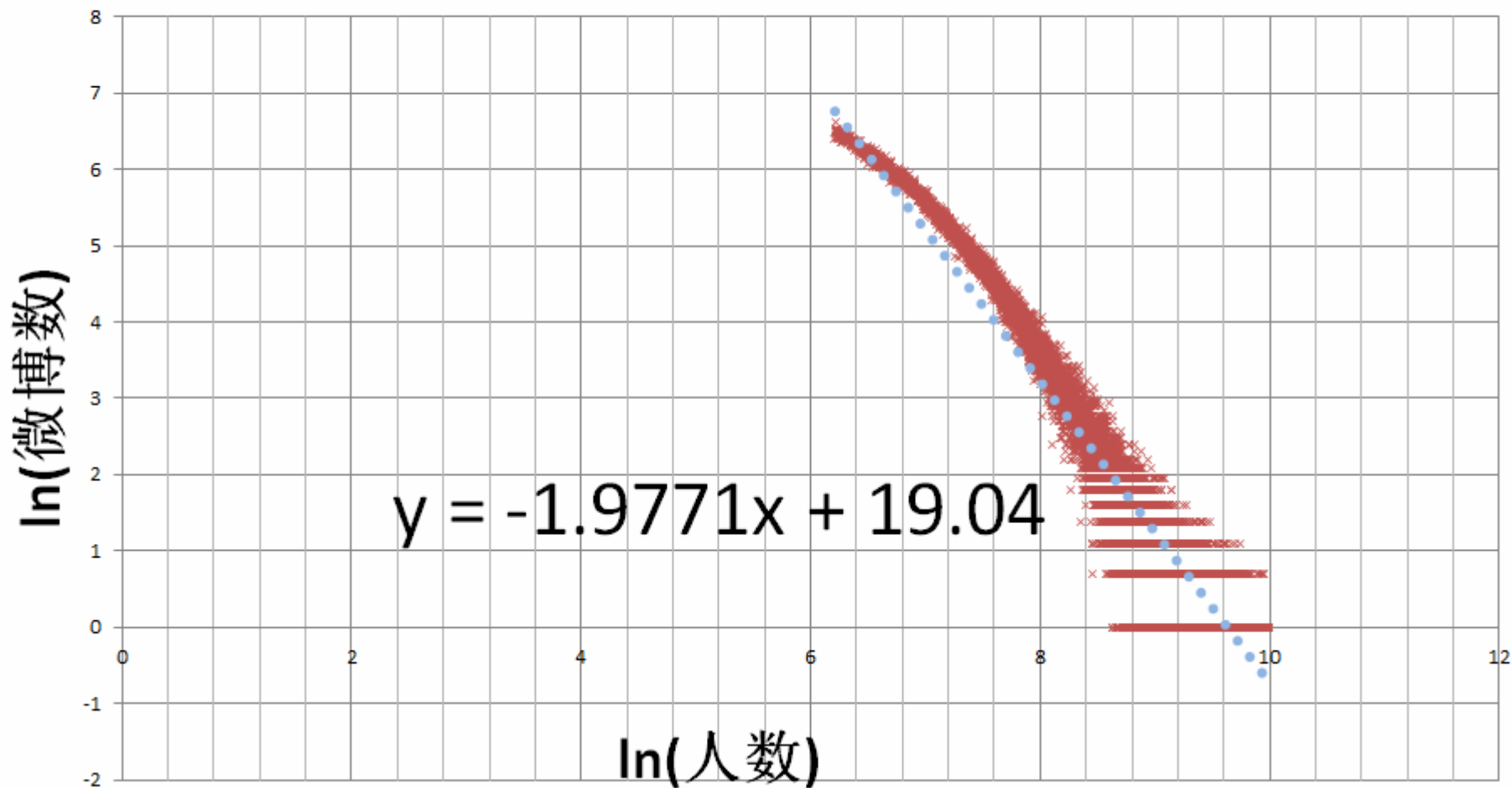
# 教育/年龄挖掘

➤ 662,565 登记了教育信息, 占总人数的3.8%;  
其中551286 大学毕业或在读, 83.20%。



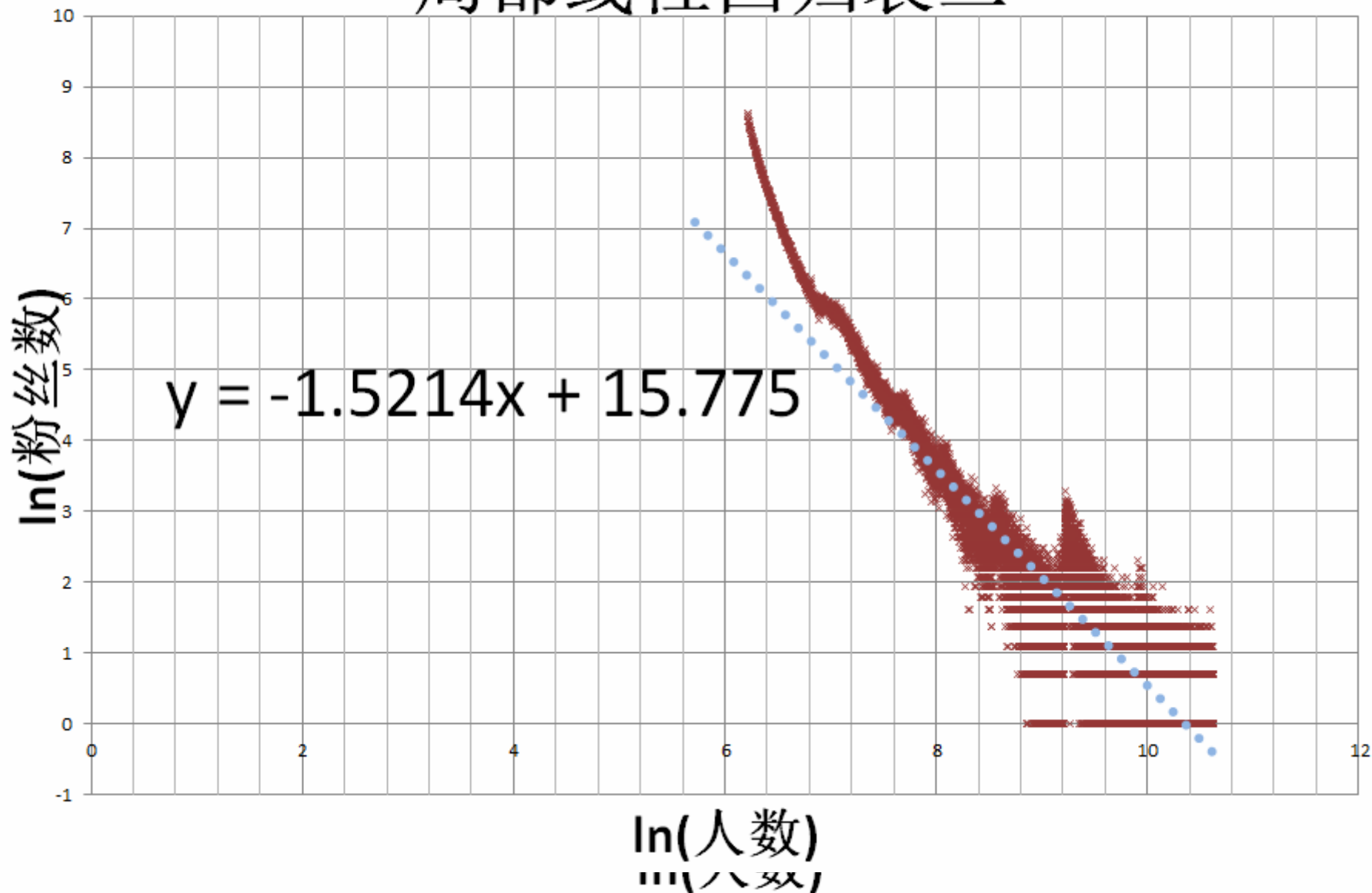
# 微博发布数规律

## 局部线性回归图二



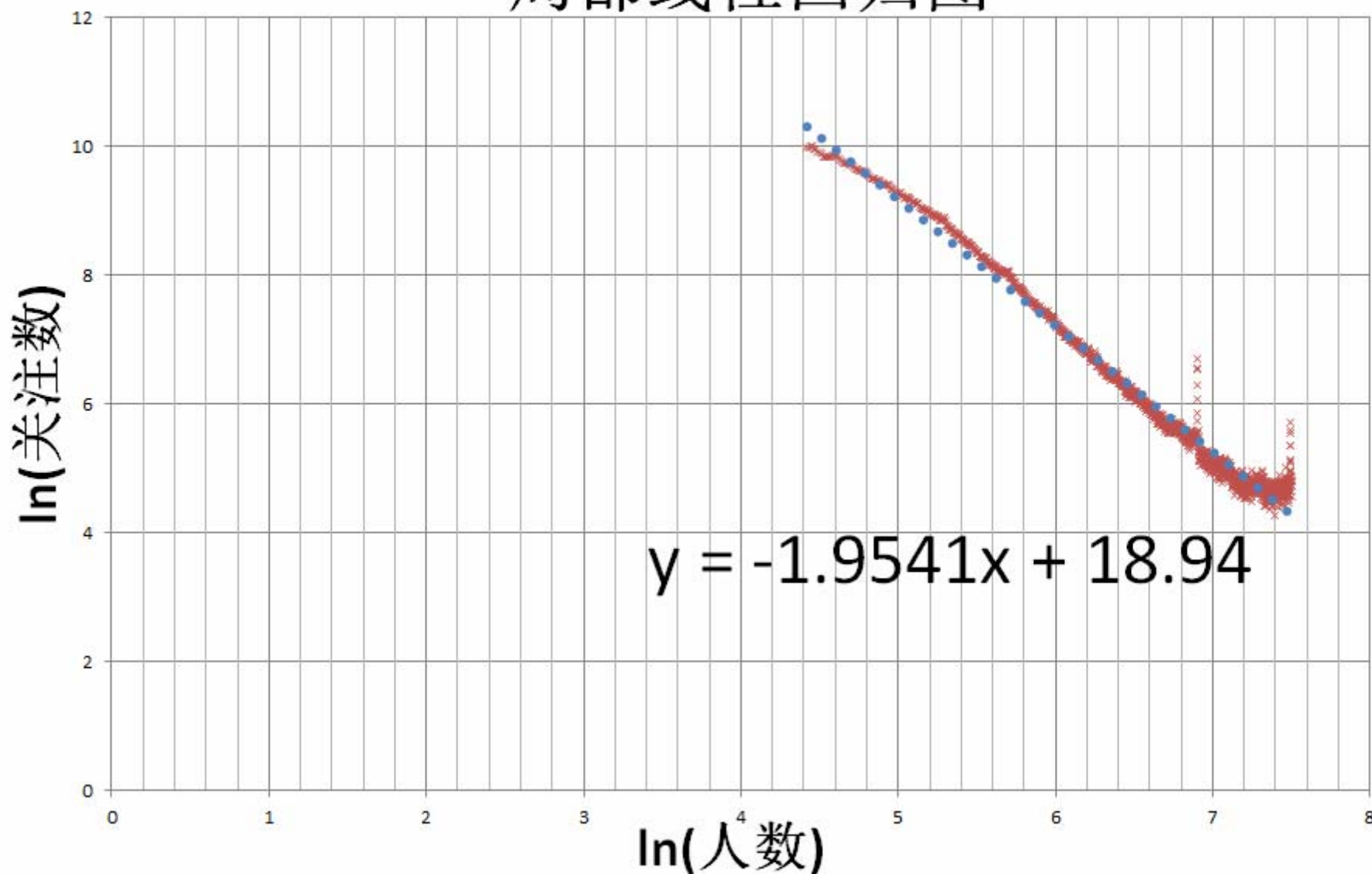
# 微博粉丝数规律

## 局部线性回归表二



# 关注数规律

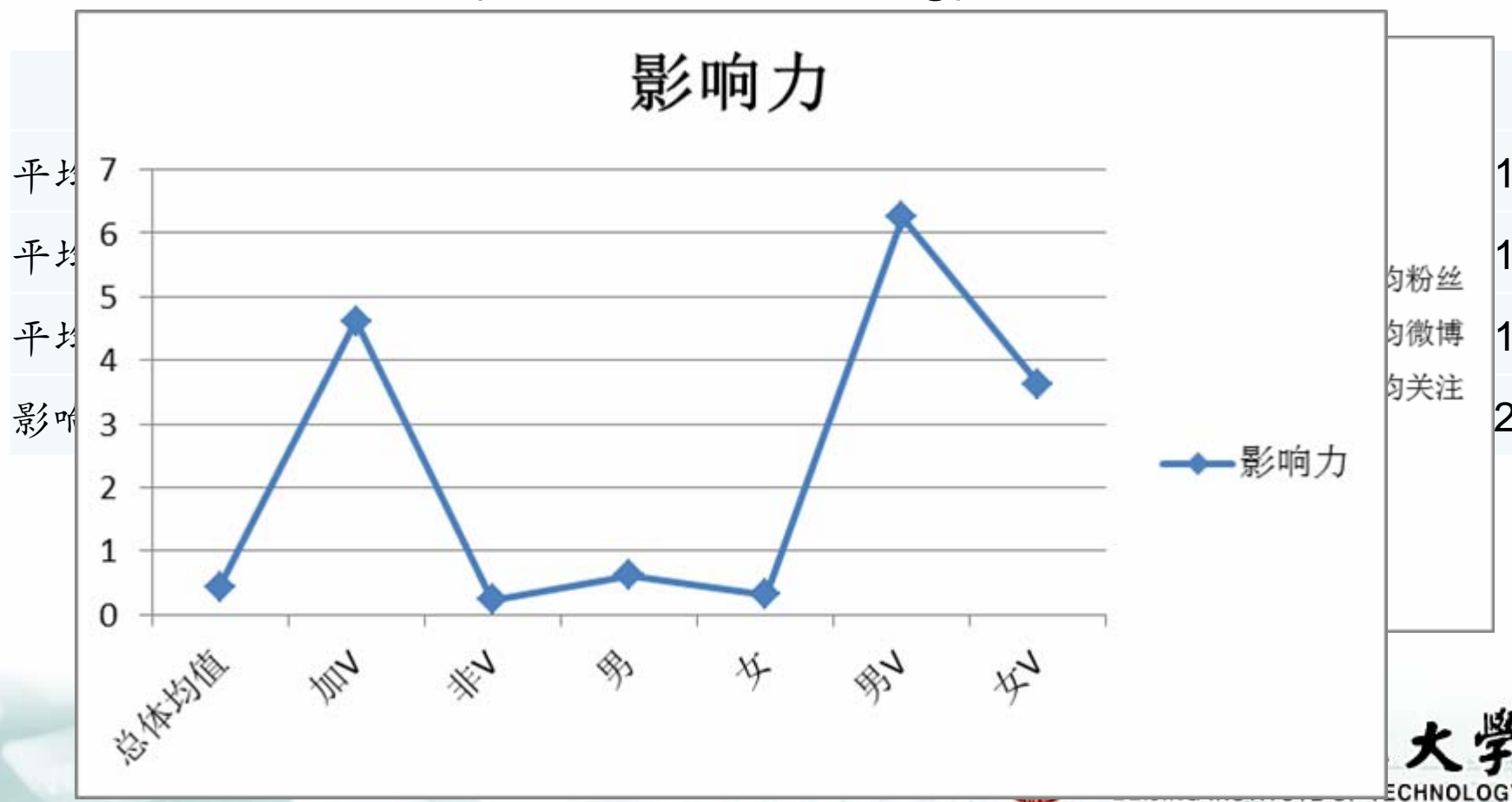
## 局部线性回归图



# 不同类型用户的影响力分析

## ➤影响力计算算法:

$$\blacksquare \text{Influence} = (\#fans - \#following) / \#tweets$$







# 自我介绍文本挖掘

词语	词频
生活	65518
自己	59370
爱	57317
喜欢	38479
关注	30909
世界	29169
人生	29126
快乐	27656
我们	27482
幸福	23417





# 微博用户特征大数据挖掘小结

- 微博数/粉丝数/关注数 为两段不同参数的幂律分布组合
- 关于男女的挖掘规律
  - 男女人数比例为45%:55%；认证比例为56%:44%；
  - 无论是所有人员还是加V用户，男性博主的影响力是女性的两倍
- 在全国范围内,地域分布密度和经济水平基本相关；在省市范围内的二级单位基本上与GDP正相关。
- 从已经登记的教育程度看，80%以上大学文化；
- 从自我介绍研判，微博以自我的生活化内容为主。





# 微观个性与行为建模

➤ 出发点：博主的一举一动一言一行，看似偶然，偶然背后有必然的行为模式与个性特征。

宏观特征大数据挖掘

➤ 已经发布新浪微博应用“微博个性热词云”：分析博主的个性，并计算不同主体个性；并研究个体兴趣的迁移变化。

微观个性与行为建模

话题与情感内容分析

➤ <http://esyrt.sinaapp.com/>



北京理工大学  
BEIJING INSTITUTE OF TECHNOLOGY

➤ 数据来源:博主发布的所有微博内容;

➤ 分析方法

- 汉语分词与词性标注: 采用博主研制的NLPIR(ICTCLAS2013);
- 利用交叉信息熵计算有代表性的关键词 $w$ , 权重 $f(w) = \sum_l -p_l \ln p_l + \sum_r -p_r \ln p_r$ ;
- 所有关键词及权重组成的向量成为博主的微观个性



# 十八大报告的关键语义分析

NLPIR汉语分词系统 (又名: ICTCLAS2013版) 张华平博士出品,新增新词发现、关键词识别与微博分词

NLPIR分词



分词



用户词典



关键词提取



指纹提取

相关介绍

坚定不移沿着中国特色社会主义道路前进 为全面建成小康社会而奋斗  
——中国共产党第十八次全国代表大会报告  
(2012年11月8日)  
胡锦涛  
同志们:  
现在,我代表第十七届中央委员会向大会作报告。  
中国共产党第十八次全国代表大会,是在我国进入全面建成小康社会决定性阶段召  
开的一次十分重要的大会。大会的主题是:高举中国特色社会主义伟大旗帜,以邓小平

打开文件 分析 清空

关键词	权重
中国特色社会主义	20.21
改革开放	11.43
科学发展观	10.36
人民生活水平	9.99
经济发展方式	9.28
社会公平正义	8.74
收入分配差距	7.92
中华民族伟大复兴	7.91
城乡发展一体化	7.91
十年	7.74
基础设施	7.73
当代中国	7.73
发展	7.52
基本公共服务	7.49
生态文明建设	7.49
和平发展	7.07
开放型经济	7.06





# 微博个性分析的交叉熵原理

➤ word=科学发展观 词频=17 交叉熵=10.43

➤ 出现的位置

■ (10316,11266,11683,12141,12144,12217,12247,12281,12302,12334,12388,12442,12513,12585,12688,13534,24612)

➤ 上文种类= 9

■ (符合(1),。(3),、(2),是(1),了(1),把(1),贯彻(1),实践(2),落实(5),)

➤ 下文种类=12

■ (, (2),的(4),为(1),。(1),要求(1),等(1),最(1),是(2),贯彻(1),同(1),活动(1),在内(1),



北京理工大学  
BEIJING INSTITUTE OF TECHNOLOGY



# 博主个性化建模：沈阳教授

武大沈阳最近的200条微博中，最热词汇是“**微博**”，总共提到了**59**次！



下面是武大沈阳在微博的最

“微博”一共出现了**59**次！

“网友”一共出现了**18**次！

“新闻”一共出现了**14**次！

“媒体”一共出现了**11**次！

“意见”一共出现了**10**次！

“兴趣”一共出现了**9**次！

“分析”一共出现了**8**次！

“领袖”一共出现了**8**次！

“公益”一共出现了**7**次！

“标题”一共出现了**7**次！

分享到微博

#微博个性热词云# 呵呵目前 @武大  
博”！#微博个性热词云#还可以查看  
来<http://esyrtsinaapp.com>试一试吧



分享到微博



北京理工大学  
BEIJING INSTITUTE OF TECHNOLOGY

# 张华平的个性化特征演化

## 2011年9月20日

亲爱的ICTCLAS张华平博士你好！

你最近的200条微博中最热的词汇是“网络”，一共出现了23次！

小窗口播放

社理工  
实验室 北京 信息 处理  
大学 网络  
安全 平台 搜索 语言

“安全”一共出现了9次！ “应用”一共出现了9次！ “学习”一共出现了9次！  
“教授”一共出现了9次！ “准备”一共出现了8次！ “团队”一共出现了8次！  
“参加”一共出现了8次！ “问题”一共出现了8次！ “研究生”一共出现了8次！  
“事件”一共出现了8次！ “检索”一共出现了8次！ “事情”一共出现了7次！  
“个人”一共出现了7次！ “共享”一共出现了7次！ “管理”一共出现了7次！  
“技术”一共出现了7次！ “值得”一共出现了7次！ “学生”一共出现了7次！  
“发布”一共出现了7次！ “系统”一共出现了7次！ “过程”一共出现了7次！  
“但愿”一共出现了6次！ “张华”一共出现了6次！ “采用”一共出现了6次！  
“年前”一共出现了6次！ “领导”一共出现了6次！ “访问”一共出现了6次！  
“自动”一共出现了6次！ “有意思”一共出现了6次！ “比较”一共出现了6次！  
“兴趣”一共出现了6次！ “精神”一共出现了6次！ “效果”一共出现了6次！  
“演讲”一共出现了6次！ “交流”一共出现了6次！ “独立”一共出现了6次！

分享到微博

#微博热词云# 呵呵我目前的微博最热词是“网络”！#微博热词云  
#还可以查看对比Ta们的微博相似度哦！快来  
<http://esyt.sinaapp.com>试一下吧！



分享到微博

切换浏览模式 IE打开 下载 100%

# 张华平的个性化特征演化

## 2012年2月25日

我的微博热词排行 看看Ta的微博热词排行 查看Ta们的微博相似度指数！

亲爱的ICTCLAS张华平博士你好！

你最近的200条微博中最热的词汇是“**微博**”，一共出现了**27**次！

你微博的热词排名：

“微博”一共出现了27次！ | “网络”一共出现了22次！ | “研究”一共出现了17次！ |  
“挖掘”一共出现了17次！ | “实验室”一共出现了14次！ | “研究生”一共出现了12次！ |  
“搜索”一共出现了11次！ | “博士”一共出现了11次！ | “技术”一共出现了9次！ |  
“安全”一共出现了9次！ | “专家”一共出现了9次！ | “发布”一共出现了8次！ |  
“访问”一共出现了8次！ | “应用”一共出现了8次！ | “舆情”一共出现了8次！ |  
“计算机”一共出现了8次！ | “报告”一共出现了8次！ | “北理工”一共出现了7次！ |  
“内容”一共出现了7次！ | “小时”一共出现了7次！ | “同学”一共出现了7次！ |  
“分析”一共出现了7次！ | “专业”一共出现了6次！ | “时间”一共出现了6次！ |  
“nlp”一共出现了6次！ | “蒙牛”一共出现了6次！ | “地址”一共出现了6次！ |  
“教授”一共出现了6次！ | “工作”一共出现了6次！ | “计算”一共出现了6次！ |

分享到微博

呵呵我目前的微博最热词是“微博”！#微博个性热词云#还可以查看对比Ta们的微博相似度哦！快来<http://esyt.sinaapp.com>试一试吧！



分享到微博



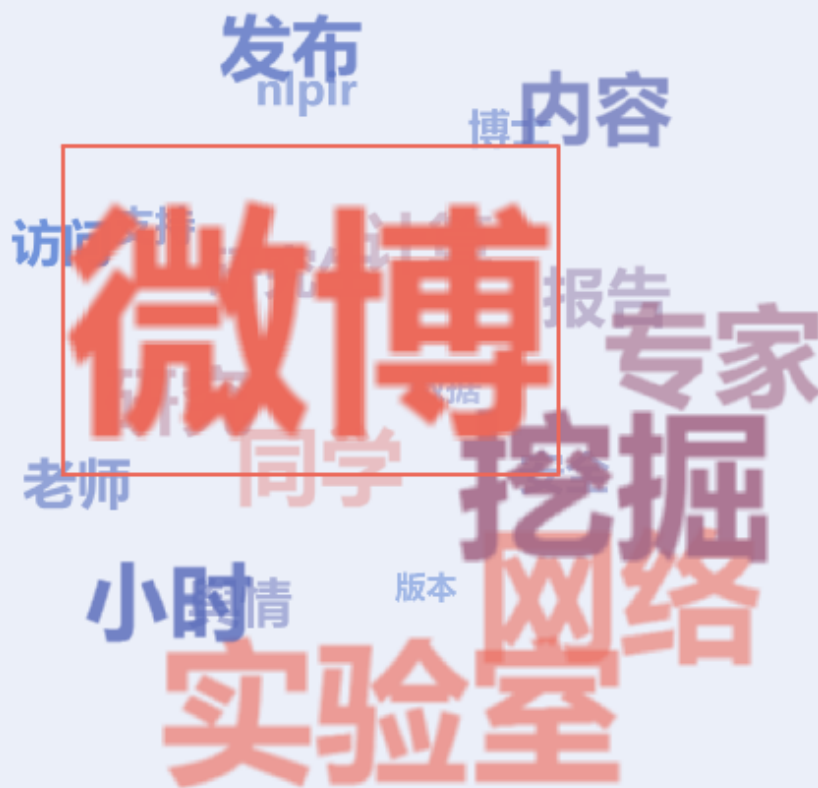
北京理工大学  
BEIJING INSTITUTE OF TECHNOLOGY

# 张华平的个性化特征演化

## 2013年4月23日

亲爱的ICTCLAS张华平博士你好！

你最近的200条微博中最热的词汇是“**微博**”，一共出现了**25**次！



你微博的热词排名：

- “微博”一共出现了**25**次！
- “同学”一共出现了**11**次！
- “研究”一共出现了**10**次！
- “计算”一共出现了**9**次！
- “舆情”一共出现了**8**次！
- “支持”一共出现了**7**次！
- “数据”一共出现了**7**次！
- “老师”一共出现了**7**次！
- “准备”一共出现了**6**次！
- “2012”一共出现了**6**次！

分享到微博

呵呵！我目前的微博最热词是“微博”  
看对比Ta们的微博相似度哦！快来ht  
吧！



分享到微博

大学

BEIJING INSTITUTE OF TECHNOLOGY

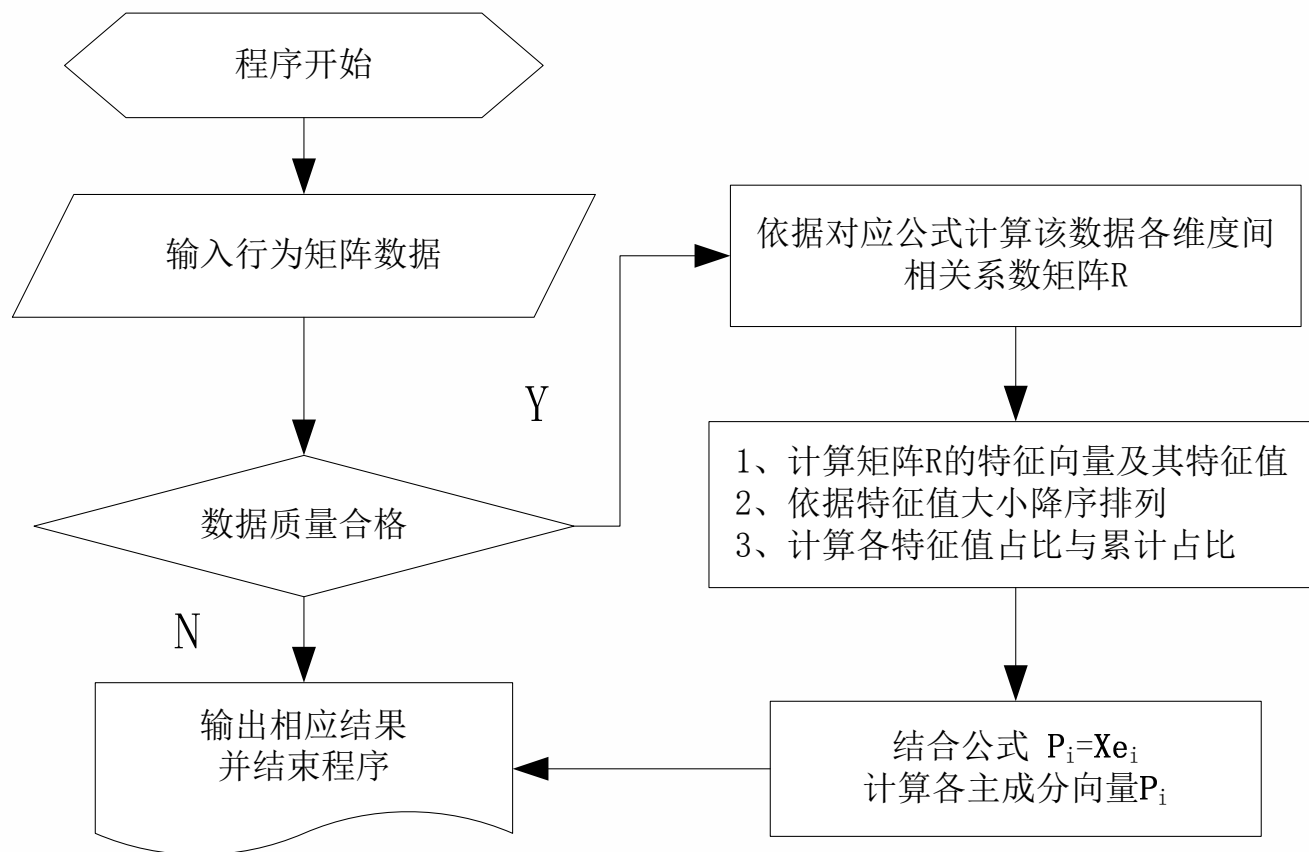


# 微博博主微观行为建模

日期\时段	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	总计
2011/1/1	0	2	0	2	0	0	1	0	1	5	0	0	0	0	0	0	0	1	0	0	0	0	0	0	13
2011/1/2	0	3	4	0	0	2	0	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	14
2011/1/3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	3	0	1	5	
2011/1/4	1	1	0	0	2	0	0	0	0	0	0	0	0	0	1	1	0	1	0	0	0	0	0	0	9
2011/1/5	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	3	2	1	0	0	0	0	0	0	16
2011/1/6	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	9
2011/1/7	0	0	0	0	2	0	2	0	0	0	2	0	0	2	1	0	1	0	0	0	0	0	0	0	10
2011/1/8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	3	1	2	0	1	0	0	0	8
2011/1/9	0	0	0	0	0	0	0	0	0	0	4	2	1	1	0	0	4	0	0	0	0	0	1	0	13
2011/1/10	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	3	0	0	0	0	0	2	3	0	9
2011/1/11	0	0	0	0	0	0	0	0	0	3	1	0	0	3	25	0	0	2	0	0	0	0	0	0	34
2011/1/12	0	0	0	0	0	0	0	0	1	1	0	0	3	0	0	0	3	1	3	0	0	0	0	2	14
2011/1/13	2	0	0	0	0	0	0	0	0	0	0	2	0	0	2	1	0	0	1	1	1	0	0	0	10
2011/1/14	0	0	0	0	0	0	0	0	1	1	2	0	1	1	1	1	0	0	3	0	0	2	1	0	14
2011/1/15	0	0	0	0	0	0	0	1	1	1	1	2	0	0	0	1	4	2	0	0	0	1	0	1	15
2011/1/16	0	0	0	0	0	0	0	0	1	1	3	2	0	0	0	2	2	2	2	0	0	1	4	1	21
2011/1/17	0	0	0	0	0	0	0	1	1	1	0	1	0	1	1	4	3	4	1	2	2	1	0	0	23
2011/10/8	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	3
2011/10/9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
2011/10/10	0	0	0	0	0	0	0	0	0	0	0	0	1	5	0	0	0	1	1	0	0	0	0	1	9
2011/10/11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	1	0	1	0	0	0	5
2011/10/12	0	0	0	0	0	0	0	2	0	0	0	4	1	3	0	0	1	1	3	1	2	0	0	0	18
2011/10/13	0	0	0	0	0	0	2	1	0	1	0	0	0	0	2	0	1	0	3	1	2	1	1	0	15
2011/10/14	0	0	0	0	0	0	0	1	0	0	2	5	0	0	0	0	3	1	1	0	0	0	0	0	13
2011/10/15	0	0	0	0	0	0	0	0	0	0	0	5	0	1	1	1	1	1	3	0	0	0	0	0	13
2011/10/16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	2
2011/10/17	0	0	0	0	0	0	0	0	2	2	0	0	1	1	1	0	1	0	0	0	0	0	0	0	8
原始数量总计	46	28	14	12	12	12	48	126	190	186	196	254	171	163	266	222	233	244	258	186	142	161	207	145	3522
LOG2处理总计	36	16	10	10	9	10	40	105	150	150	136	176	130	126	162	156	166	166	191	148	118	126	150	109	2597
布尔处理总计	27	9	7	7	6	8	31	82	109	113	92	111	94	90	111	107	115	109	128	108	90	92	101	77	1824

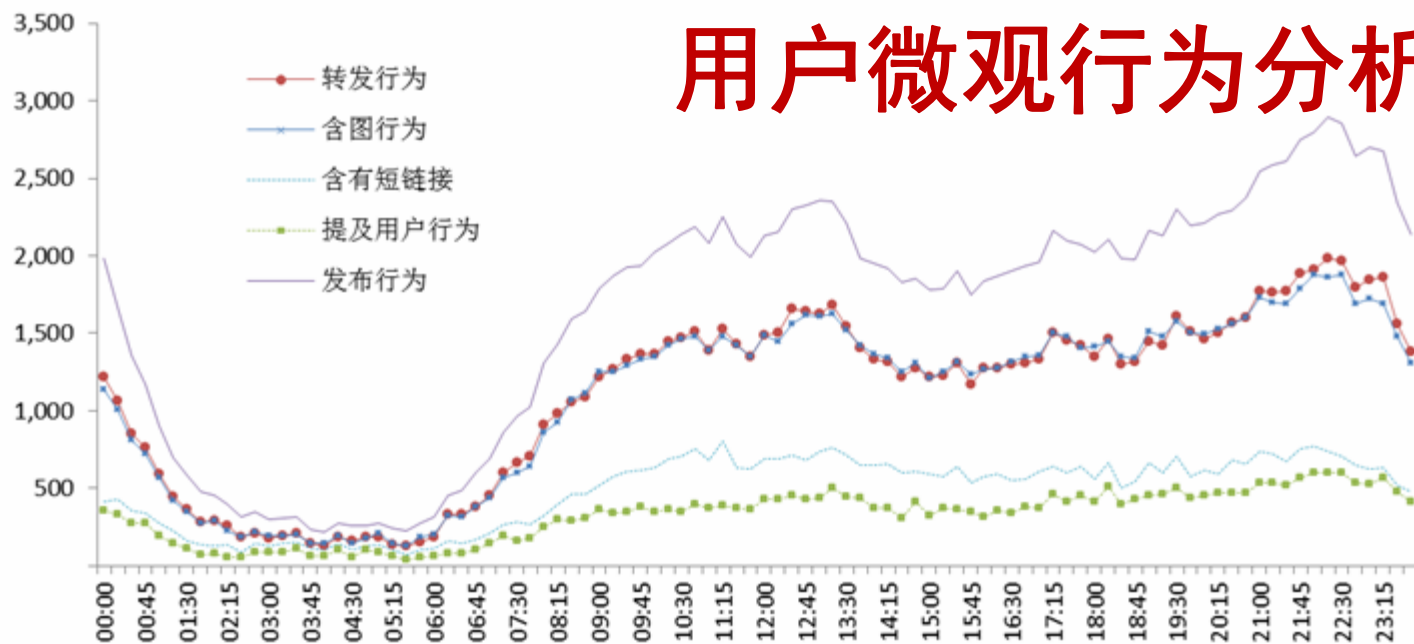
焦点定位

# 行为矩阵分析系统流程

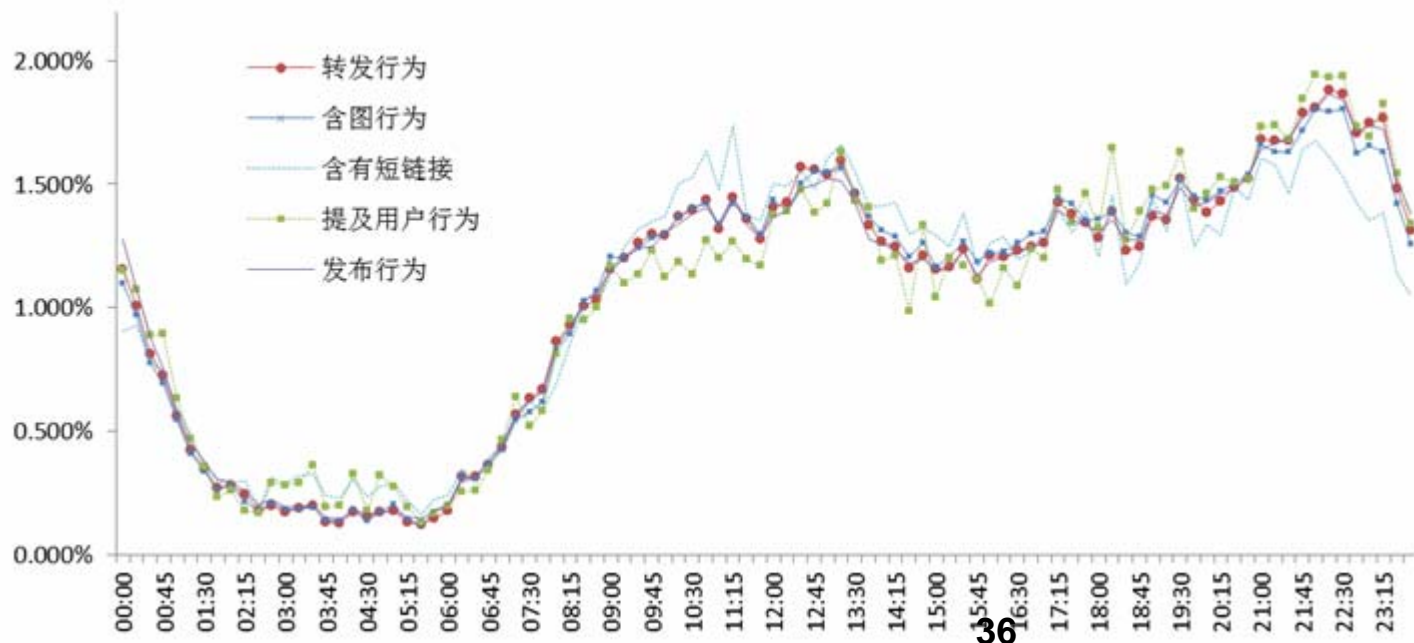


# 用户微观行为分析

第一主成分  
(原始数据)

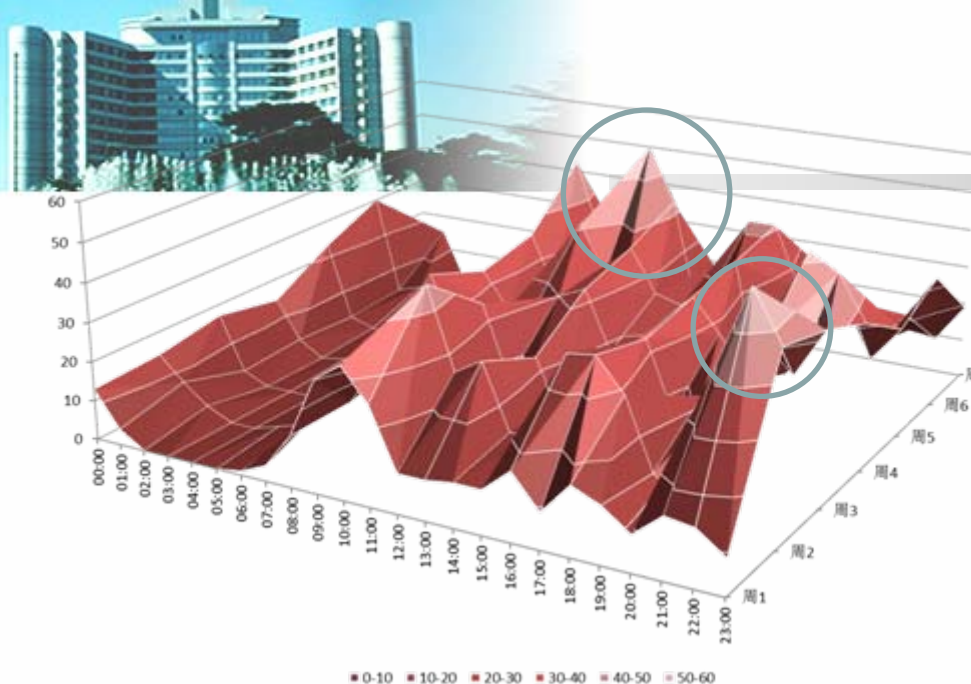


第一主成分  
(归一化)



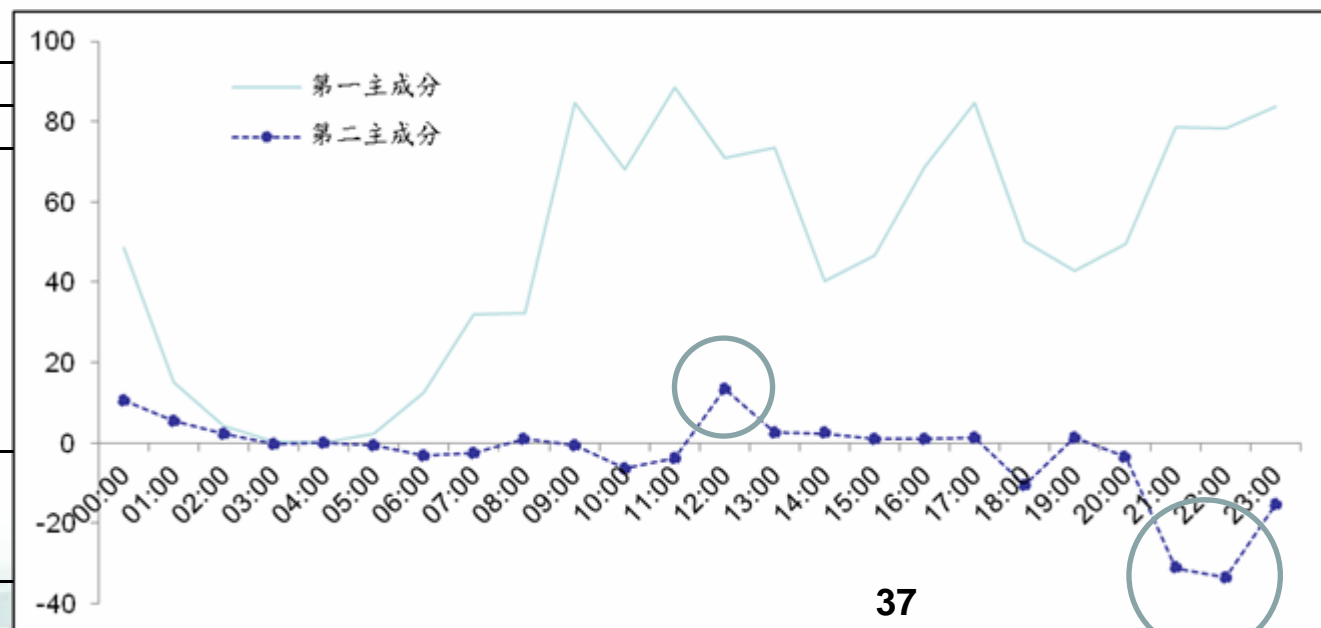


# 用户微观行为分析



$$R = \begin{bmatrix} 1.0000 & 0.6056 & 0.6291 & 0.5463 & 0.6231 & 0.6588 & 0.4417 \\ 0.6056 & 1.0000 & 0.7755 & 0.8429 & 0.6985 & 0.6368 & 0.5791 \\ 0.6291 & 0.7755 & 1.0000 & 0.8411 & 0.7931 & 0.6548 & 0.5913 \\ 0.5463 & 0.8429 & 0.8411 & 1.0000 & 0.7569 & 0.6409 & 0.5897 \\ 0.6231 & 0.6985 & 0.7931 & 0.7569 & 1.0000 & 0.8406 & 0.6914 \\ 0.6588 & 0.6368 & 0.6548 & 0.6409 & 0.8406 & 1.0000 & 0.7125 \\ 0.4417 & 0.5791 & 0.5913 & 0.5897 & 0.6914 & 0.7125 & 1.0000 \end{bmatrix}$$

	v1	v2
周1	0.3332	-0.0602
周2	0.3853	-0.3946
周3	0.3970	-0.3143
周4	0.3925	-0.3908
周5	0.4054	0.1831
周6	0.3840	0.4498
周7	0.3421	0.5945
特征值	5.0649	0.6225
占比	72.36%	8.89%
累计占比	72.36%	81.25%



# 微博博主行为模式挖掘

$$\begin{bmatrix} 1 & \text{Corr}(X_1, X_2) & \cdots & \text{Corr}(X_1, X_n) \\ \text{Corr}(X_2, X_1) & 1 & \cdots & \text{Corr}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Corr}(X_n, X_1) & \text{Corr}(X_n, X_2) & \cdots & 1 \end{bmatrix}$$

$$GM_j = \sqrt[6]{\prod_{i=1}^7 |a_{ij}|} \quad AM_j = \frac{\sum_{i=1}^7 |a_{ij}| - 1}{6}$$

相关系数矩阵	周一	周二	周三	周四	周五	周六	周日
周一	1	0.667969724	0.742039339	0.724229458	0.739878506	0.756160482	0.522685238
周二	0.667969724	1	0.855389999	0.79381239	0.850451272	0.791522972	0.662471259
周三	0.742039339	0.855389999	1	0.785204945	0.843321875	0.798761405	0.593729684
周四	0.724229458	0.79381239	0.785204945	1	0.840632355	0.845562426	0.63969534
周五	0.739878506	0.850451272	0.843321875	0.840632355	1	0.870138942	0.724187086
周六	0.756160482	0.791522972	0.798761405	0.845562426	0.870138942	1	0.728669064
周日	0.522685238	0.662471259	0.593729684	0.63969534	0.724187086	0.728669064	1
几何平均差异率	0.686824459	0.76616058	0.764297116	0.76803971	0.809359464	0.797002832	0.641049731
算术平均差异率	0.692160458	0.770269603	0.769741208	0.771522819	0.811435006	0.798469215	0.645239612

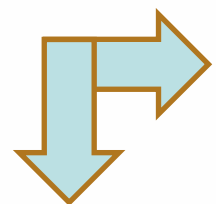
仅从作息规律而言，  
周一、周日为特殊日

- 加权求和？
- AHP？
- 向量空间的欧氏距离？
- .....

周几\属性	原创率	含图片	微博个数	几何平均差异率
1	37.78%	50.88%	11.27%	68.68%
2	33.88%	55.98%	15.67%	76.62%
3	37.54%	53.04%	17.77%	76.43%
4	36.24%	52.48%	14.34%	76.80%
5	41.67%	54.17%	15.67%	80.94%
6	46.20%	41.68%	13.83%	79.70%
7	46.40%	42.93%	11.44%	64.10%



# 微博行为模式比较

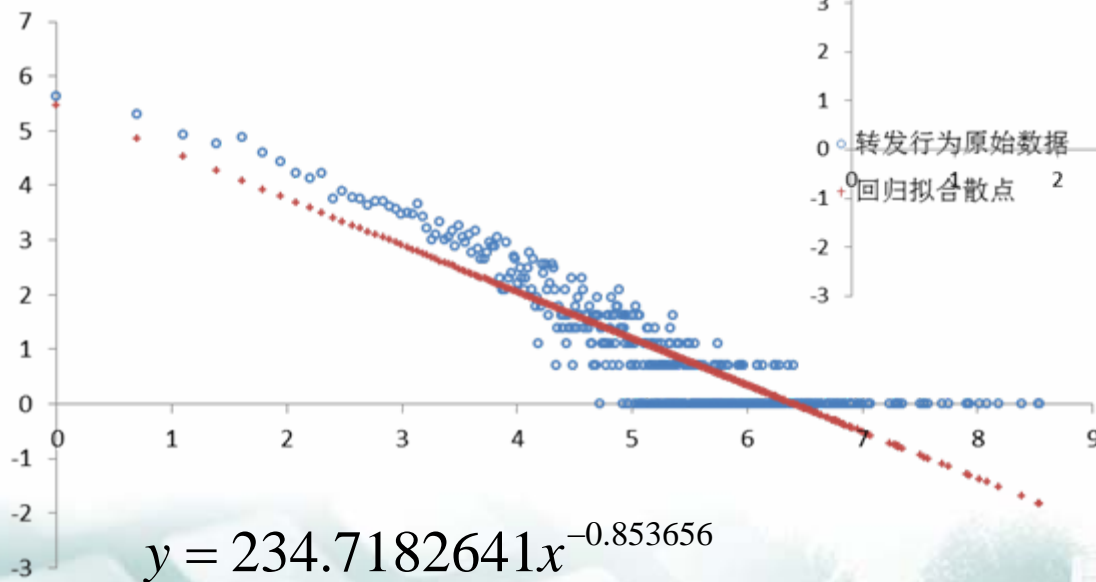
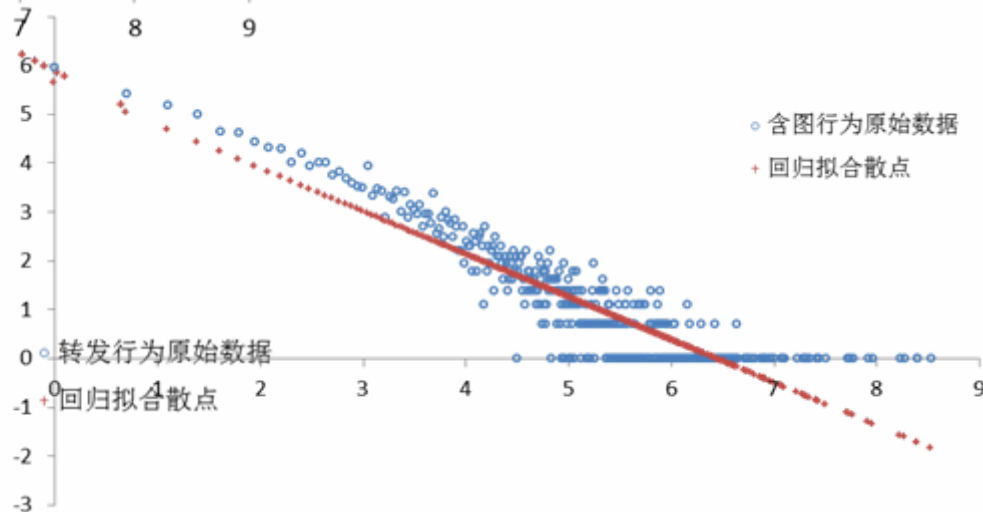
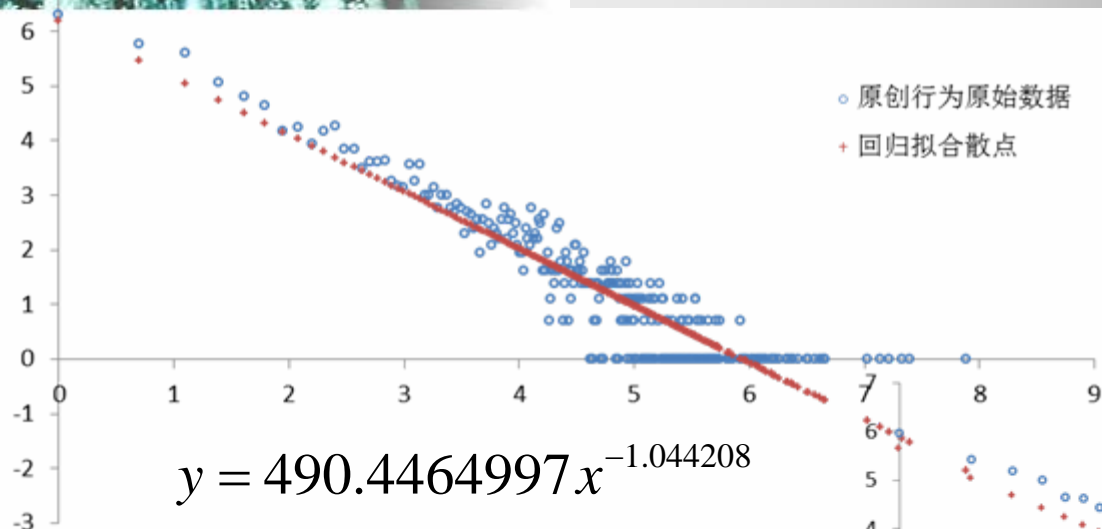


	张华平	任志强	潘石屹	张鸣	白硕	林伯强	张栋	方文山	刘强东
张华平	1								
任志强	0.447339	1							
潘石屹	0.746915	0.760761	1						
张鸣	0.84744	0.612968	0.818428	1					
白硕	0.698806	0.644066	0.81019	0.704533	1				
林伯强	0.603462	0.343865	0.602498	0.813252	0.482863	1			
张栋	0.773073	0.465831	0.758745	0.765128	0.843614	0.700826	1		
方文山	0.073967	-0.02963	0.191023	-0.06105	-0.1434	-0.25742	-0.21359	1	
刘强东	0.759182	0.221129	0.647998	0.716517	0.67937	0.661019	0.749052	0.010954	1



# 微博用户行为转换的建模

(5000人数据)



$$y = 283.091302x^{-0.876060}$$



北京理工大学  
BEIJING INSTITUTE OF TECHNOLOGY



# 微博用户行为转换的建模

本课题提出了一种基于二进制的状态值赋值法：

$$s_k = \sum_{i=1}^n f_i \times 2^{i-1}$$

二进制法计算状态值示例

行为 序号	f4 包含短链接	f3 提及其他用户	f2 包含图片	f1 转发	十进制 状态值	二进制 状态值
b01	0	0	0	0	0	0000
b02	1	0	1	0	10	1010
b03	1	0	1	0	10	1010
b04	1	0	1	0	10	1010
b05	0	0	1	1	3	0011
b06	0	0	0	0	0	0000
b07	0	0	0	0	0	0000
b08	0	0	1	1	3	0011
b09	0	0	1	0	2	0010
b10	0	0	1	0	2	0010



# 微博用户行为转换的建模

针对于该群体数据集：

用户行为演化模式频率表

From\To	s0000	s0001	s0010	s0011	s0100	s0101	s0110	s0111	s1000	s1001	s1010	s1011	s1100	s1101	s1110	s1111
s0000	41.91%	7.81%	8.91%	19.24%	0.76%	1.56%	0.43%	5.22%	2.84%	1.61%	2.85%	4.57%	0.11%	0.46%	0.23%	1.50%
s0001	13.24%	17.00%	4.05%	29.29%	0.50%	5.98%	0.43%	10.32%	1.62%	3.39%	1.96%	7.77%	0.06%	1.05%	0.17%	3.15%
s0010	14.69%	3.58%	39.96%	18.69%	0.49%	0.84%	1.19%	5.00%	2.09%	1.26%	3.87%	4.53%	0.12%	0.60%	0.17%	2.91%
s0011	10.13%	6.30%	4.83%	48.40%	0.33%	1.61%	0.33%	9.80%	1.78%	2.88%	2.02%	8.37%	0.08%	0.69%	0.15%	2.31%
s0100	20.55%	12.29%	8.04%	20.89%	6.34%	3.96%	5.21%	9.23%	1.70%	2.04%	1.87%	4.53%	0.45%	0.45%	0.40%	2.04%
s0101	11.23%	13.09%	3.49%	21.64%	0.43%	15.29%	0.44%	15.76%	1.55%	2.63%	1.78%	5.61%	0.14%	1.79%	0.13%	5.02%
s0110	12.34%	5.60%	19.34%	23.14%	1.70%	1.45%	12.94%	10.39%	1.65%	1.55%	2.55%	4.05%	0.25%	0.95%	0.60%	1.50%
s0111	8.94%	7.46%	4.05%	28.93%	0.52%	3.90%	0.50%	26.08%	1.76%	2.33%	1.88%	6.67%	0.15%	1.39%	0.24%	5.22%
s1000	11.99%	3.34%	4.72%	13.74%	0.23%	0.95%	0.21%	4.28%	38.88%	2.86%	10.91%	4.89%	0.51%	0.61%	0.33%	1.53%
s1001	5.50%	4.43%	2.23%	18.40%	0.22%	1.11%	0.20%	5.24%	1.81%	37.00%	1.57%	18.85%	0.15%	1.12%	0.13%	2.05%
s1010	9.33%	2.84%	6.60%	11.96%	0.19%	0.76%	0.34%	3.75%	7.49%	1.38%	45.80%	5.05%	0.28%	0.46%	1.20%	2.57%
s1011	7.39%	5.39%	3.44%	25.28%	0.22%	1.41%	0.25%	6.58%	1.88%	8.93%	2.31%	32.21%	0.11%	0.78%	0.17%	3.63%
s1100	8.54%	3.98%	5.15%	15.46%	0.29%	1.77%	0.88%	5.60%	10.31%	3.68%	8.84%	7.07%	20.32%	2.80%	3.39%	1.91%
s1101	8.27%	7.60%	5.33%	20.19%	0.21%	4.16%	0.32%	14.67%	2.05%	4.72%	2.24%	8.72%	0.27%	9.63%	0.19%	11.42%
s1110	9.93%	3.18%	4.78%	12.51%	0.38%	1.06%	0.61%	5.16%	3.49%	1.59%	17.89%	7.05%	1.14%	0.23%	27.60%	3.41%
s1111	3.97%	3.43%	4.07%	11.17%	0.24%	1.72%	0.16%	8.21%	0.90%	1.55%	1.80%	5.08%	0.11%	1.68%	0.11%	55.80%

分析该演化模式频率矩阵的特征值与特征向量可知，其仅存在唯一最大特征值 1，

故而方程

$$\pi = F^T \pi$$

存在唯一解，也即特征值 1 所对应的特征向量。



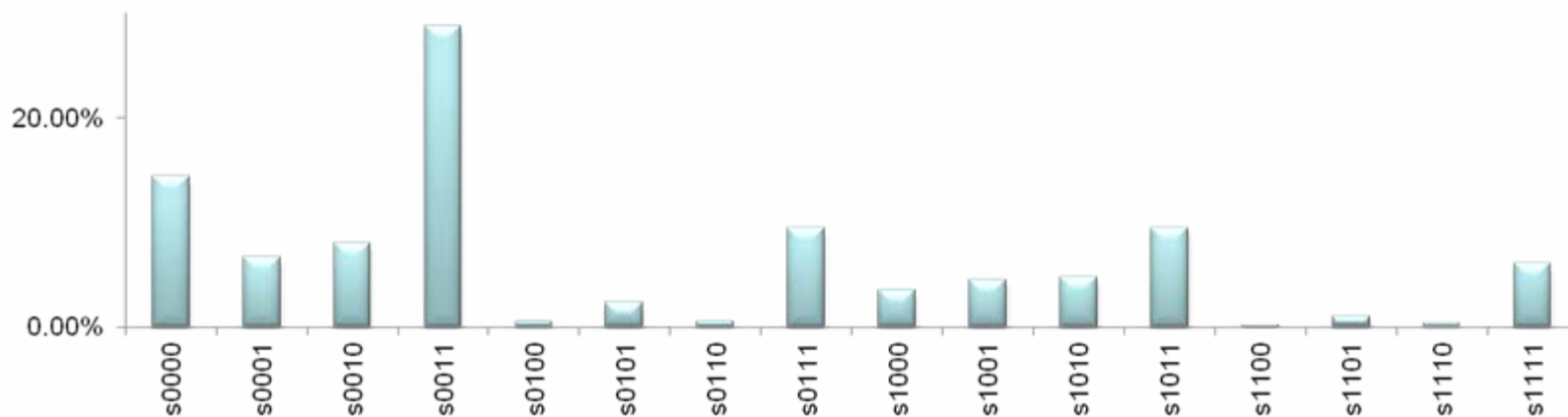


# 微博用户行为转换的建模

将该向量进行归一化处理，即可得到该演化模式的极限分布，也即稳态向量

	s0000	s0001	s0010	s0011	s0100	s0101	s0110	s0111	s1000	s1001	s1010	s1011	s1100	s1101	s1110	s1111
最大特征向量	0.3830	0.1779	0.2104	0.7634	0.0116	0.0610	0.0131	0.2498	0.0923	0.1198	0.1244	0.2516	0.0044	0.0246	0.0086	0.1608
归一化	0.1442	0.0670	0.0792	0.2873	0.0044	0.0230	0.0049	0.0940	0.0347	0.0451	0.0468	0.0947	0.0017	0.0092	0.0032	0.0605

假设微博群体发布微博的内在演化规律长期维持不变，长此以往，微博发布行为状态将最终收敛于状态







# 话题与情感内容分析

## ➤ 微博话题追踪与明码暗语发现

- 通过元搜索采集微博;
- 计算关键词义
- 综合计算词分

宏观特征大数据挖掘  
微观个性与行为建模

话题与情感内容分析

## ➤ 微博情感分析与博主情绪感知



北京理工大学  
BEIJING INSTITUTE OF TECHNOLOGY



# “明码暗语”识别

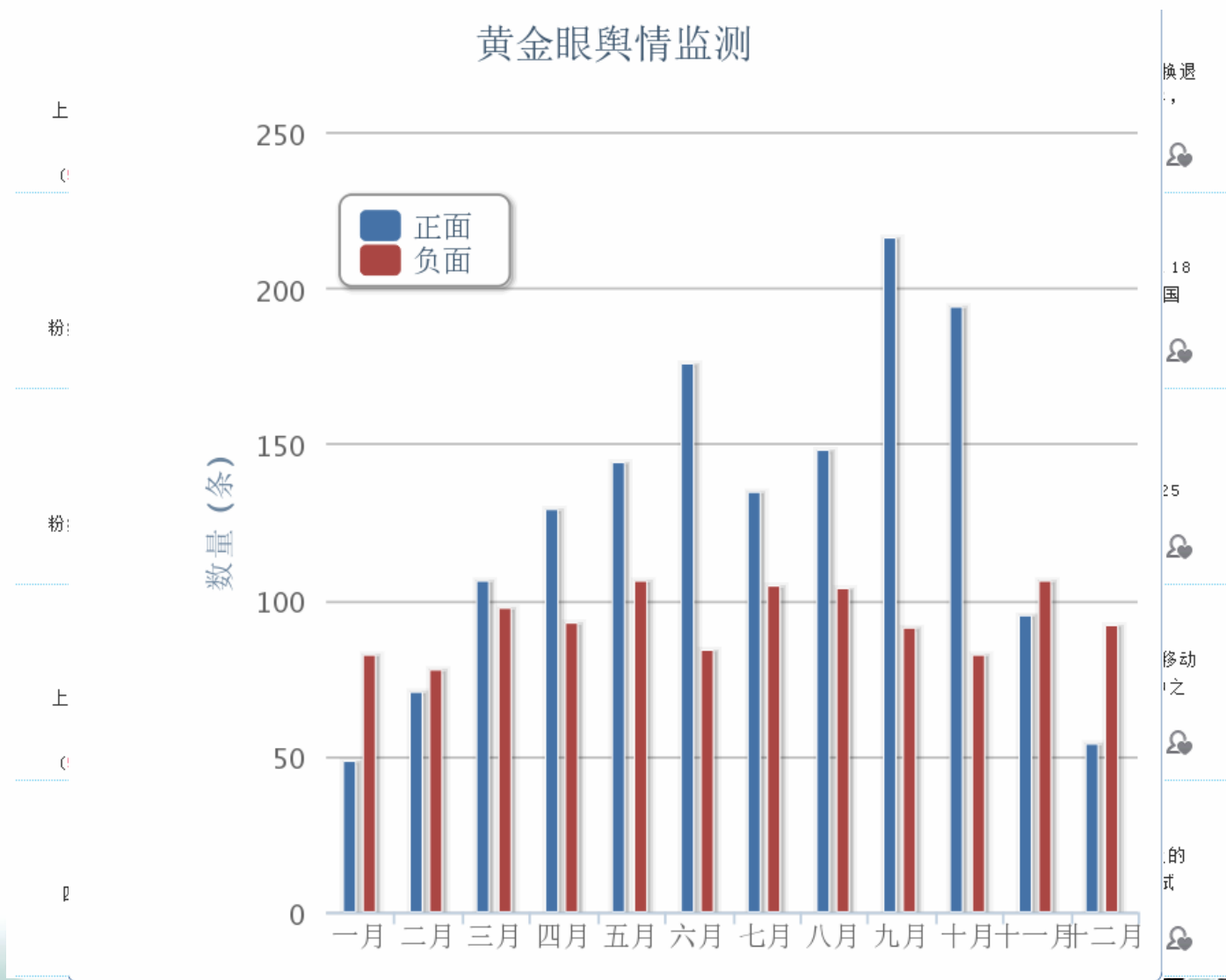


理工大学

BEIJING INSTITUTE OF TECHNOLOGY

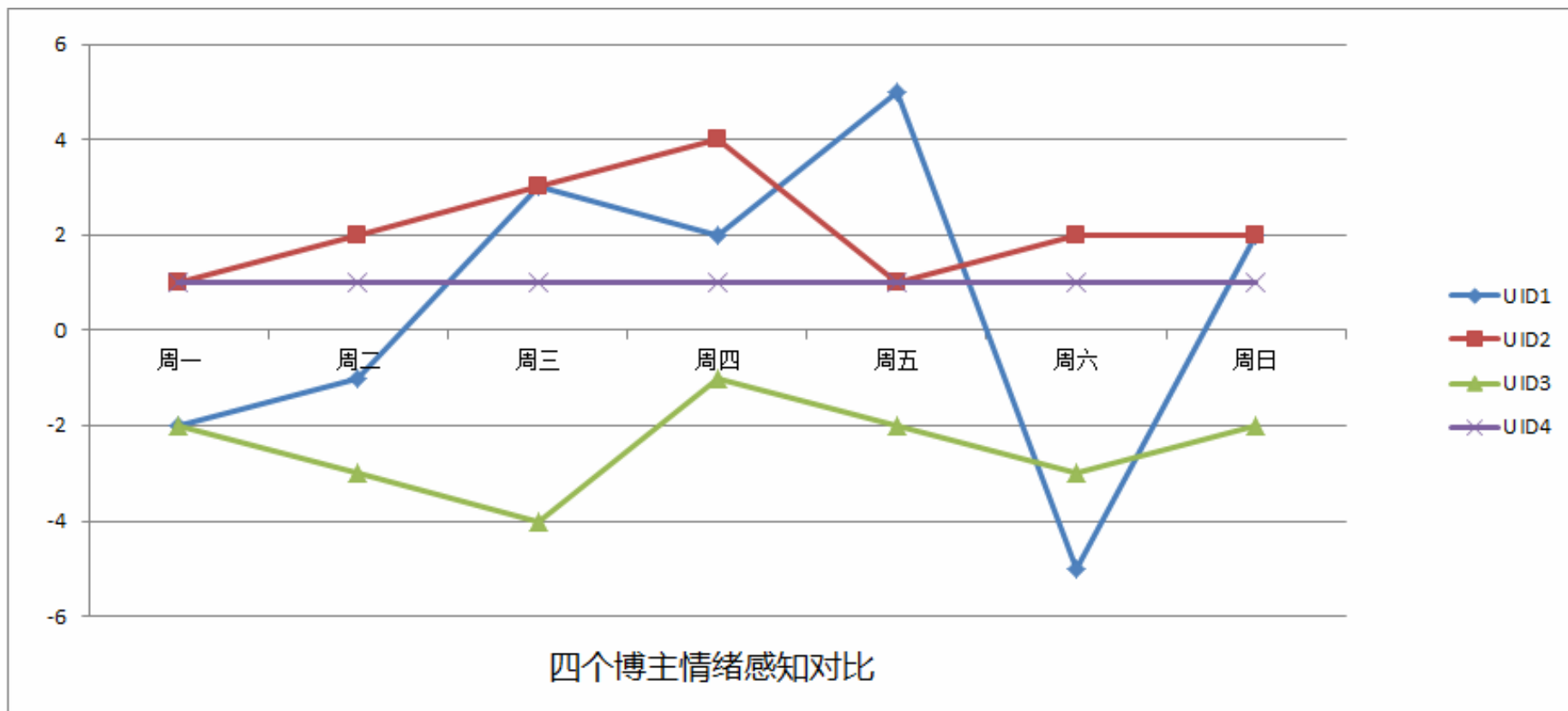
# 微博情感分析

黄金眼舆情监测





# 微博博主情绪感知





➤ 宏观：从收集的所有微博数据出发，让数据说话。挖掘各字段的相关性，实现宏观大数据挖掘；

宏观特  
征大数  
据挖掘

➤ 微观：从个人的微博内容与行为矩阵，建立个性与行为模型；

微观个  
性与行  
为建模

话题与  
情感内  
容分析

➤ 内容：为构建内容模型，对特定话题与情感进行分析。





# Thank you

相关研究由社交网络国家973课题(2013CB329606)、国家自然科学基金(61272362)支持



Contact

Email: [kevinzhang@bit.edu.cn](mailto:kevinzhang@bit.edu.cn)

Welcome to visit our homepage

<http://www.nlpir.org>

@ICTCLAS张华平博士

新浪微群: 围脖计算

微博  
计算



北京理工大学  
BEIJING INSTITUTE OF TECHNOLOGY