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

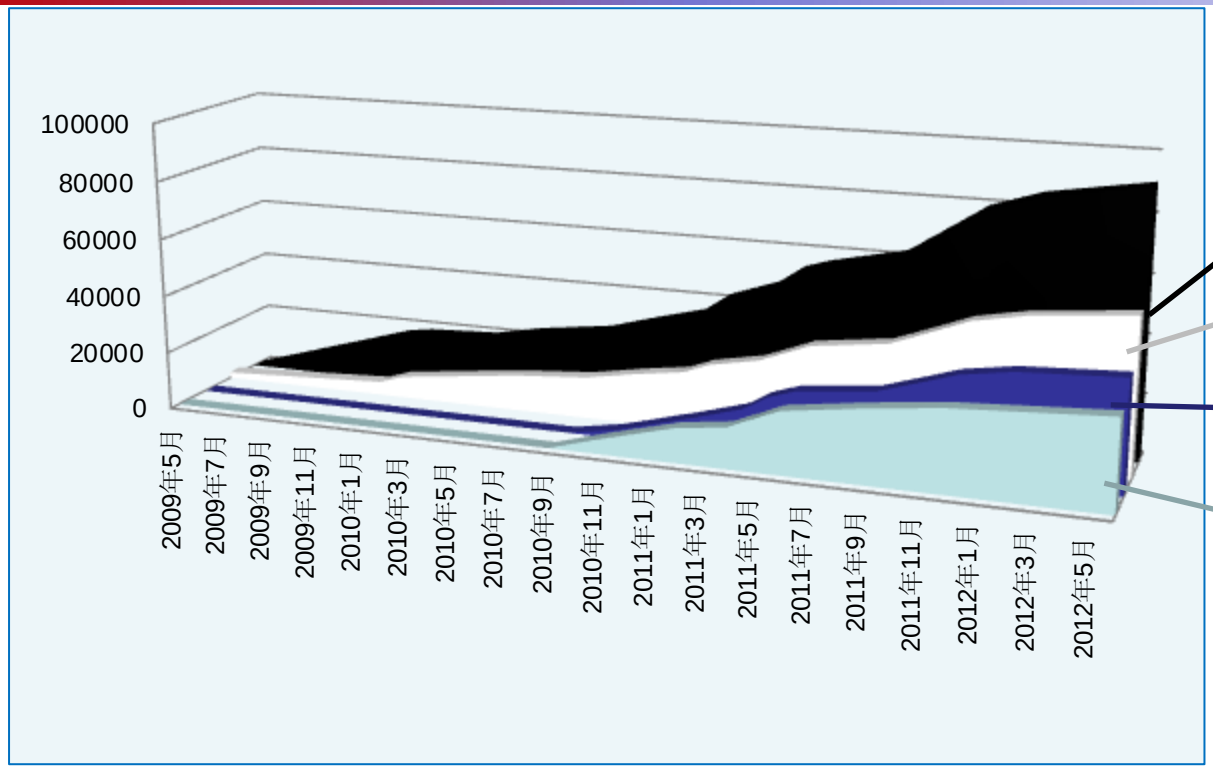
http://www.nlpir.org/

2013大数据全球技术峰会

2013/4/27

北京理工大學
BEIJING INSTITUTE OF TECHNOLOGY

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

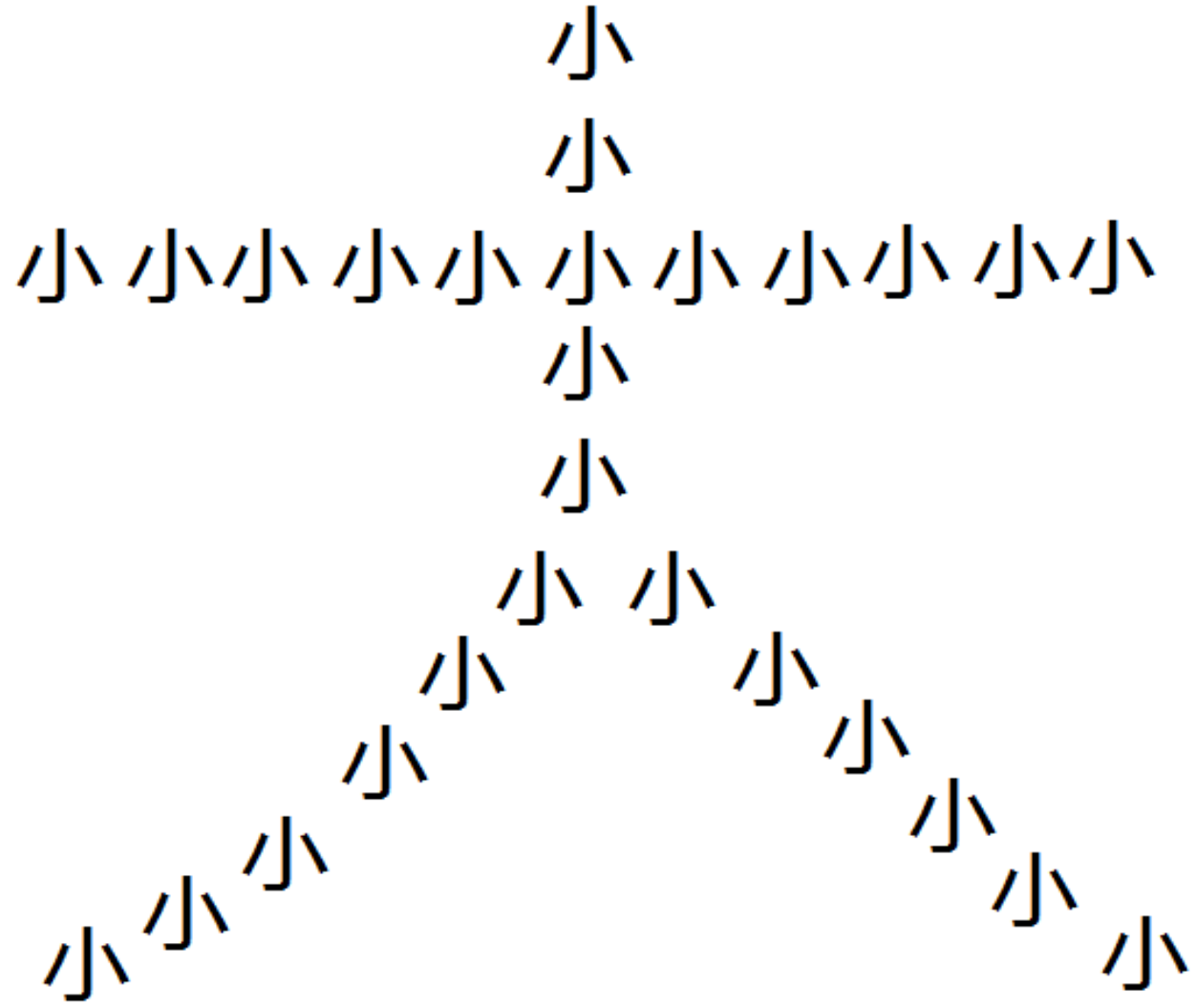
FORTUNE Money

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.

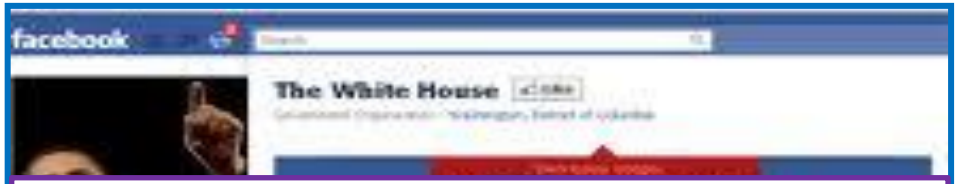
大数据 vs. 小数据



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

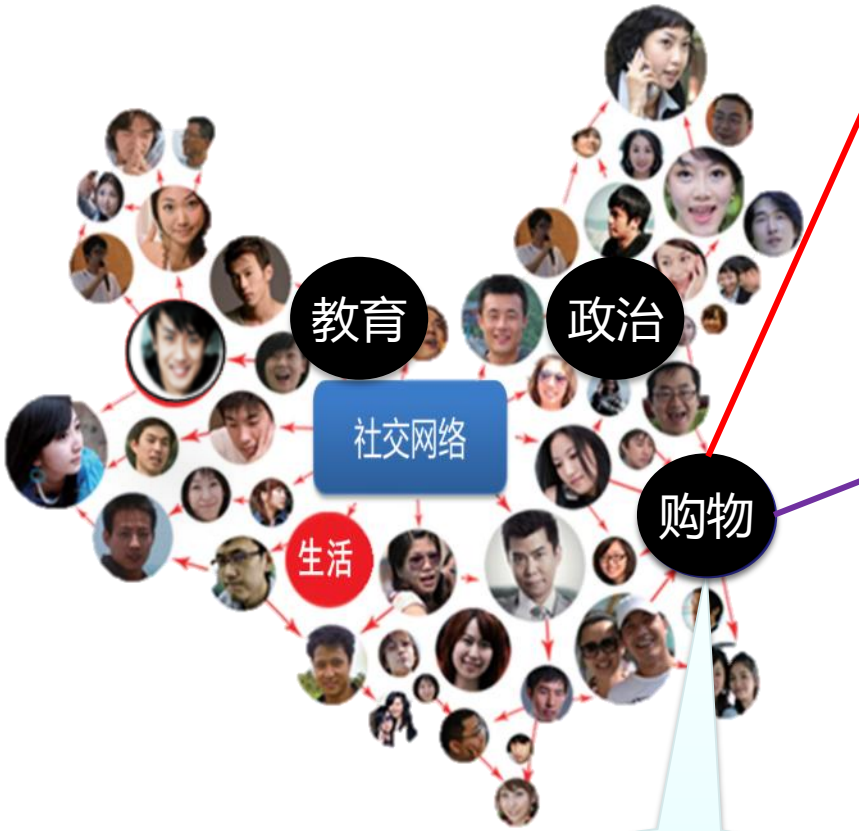


谣 言： 我国民众受微博谣言蛊惑而抢盐，影响社会稳定

在线社交网络对生活方式的影响

公开课： 超过50所美国大学在社交网络上发布公开课

在线社交网络对生活方式的影响



影响
价值观

促 销： 70%的社交网络成人活跃用户选择网上购物

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

宅生活： 网民利用社交网络可以不出家门进行交友、游戏、互动、协作

影响人生观

社交网络与商业应用

- 宏观决策：为我们提供了难得的人口显式特征与潜在特征的普查，样本=总体，实时，相对真实，最低代价；
- 微观精准：个人[illegible]研，推荐与精准营销；
- 内容理解：从语义理解真实意图，为我们提供了新的认识手段。

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/生日/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上。

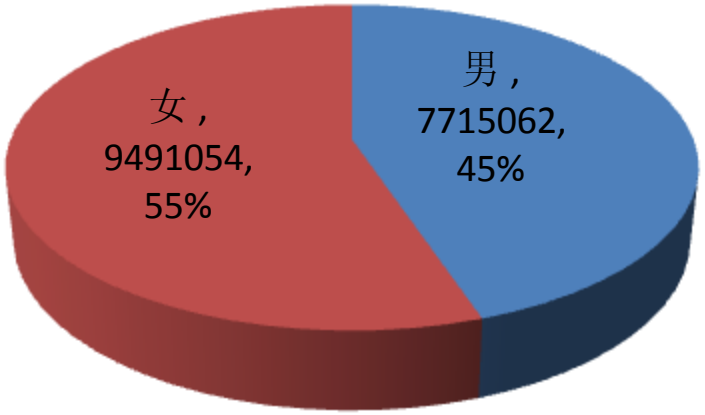
微博用户数据样本

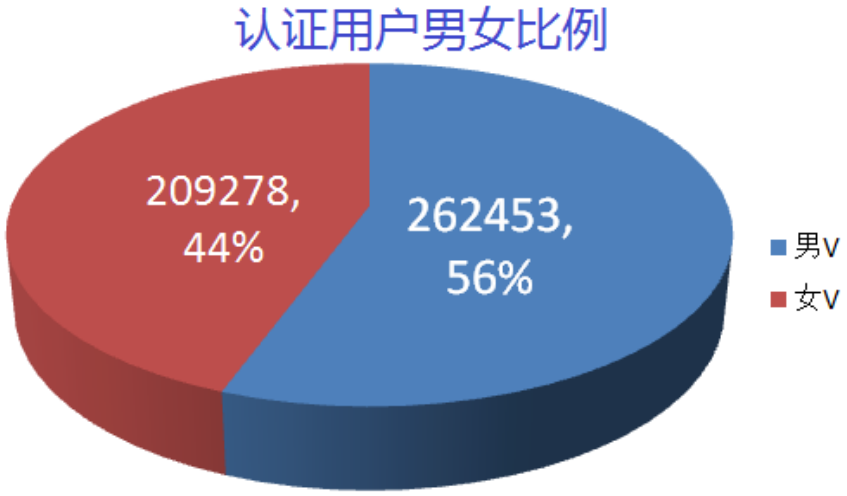
id	url	name	s..	brithday	address	fansNum	summary	wbNum	gzNum	blog	realName
10315	http://weibo.com	老军鹏	男	<NULL>	北京,海淀区	209	他还没填写个人介:	446	157	<NULL>	<NULL>
10318	http://weibo.com	王海荣	男	<NULL>	<NULL>	600	<NULL>	<NULL>	<NULL>	<NULL>	<NULL>
10362	http://weibo.com	kanglicoco	女	<NULL>	<NULL>	454	懂点人事儿的小清:	837	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	大里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htt	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 #美居	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

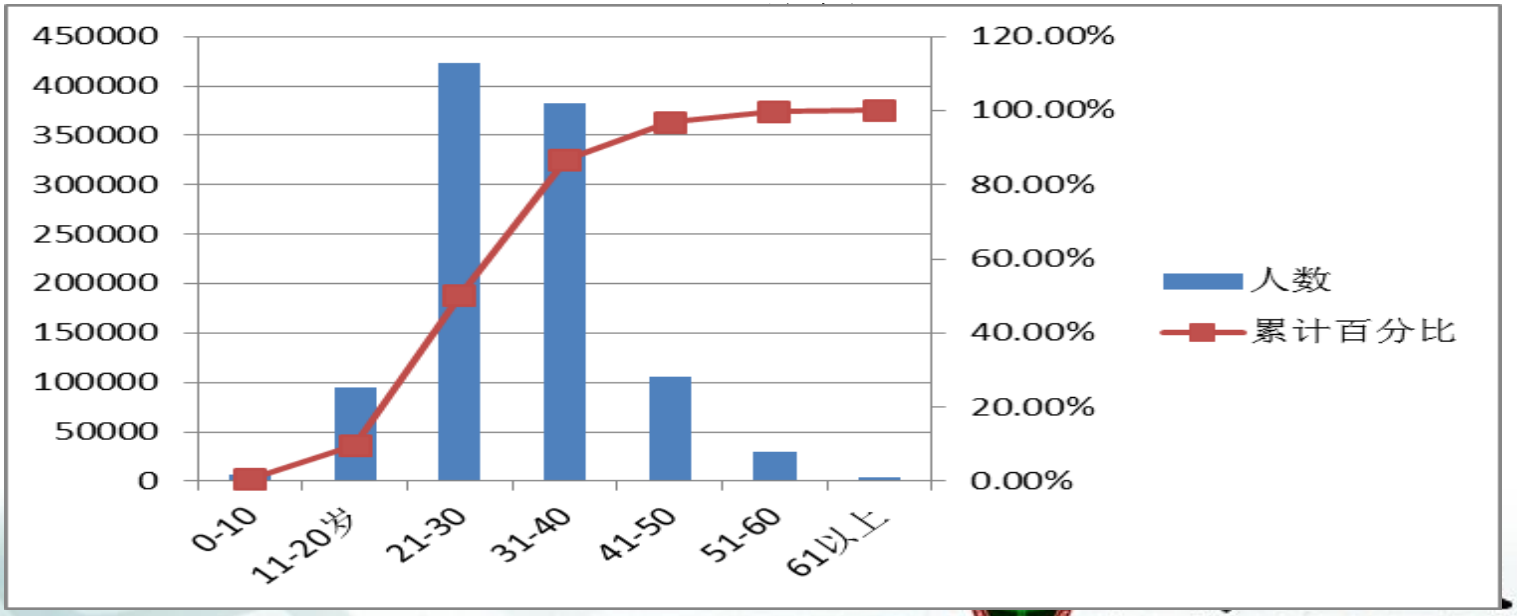
男女比例图表





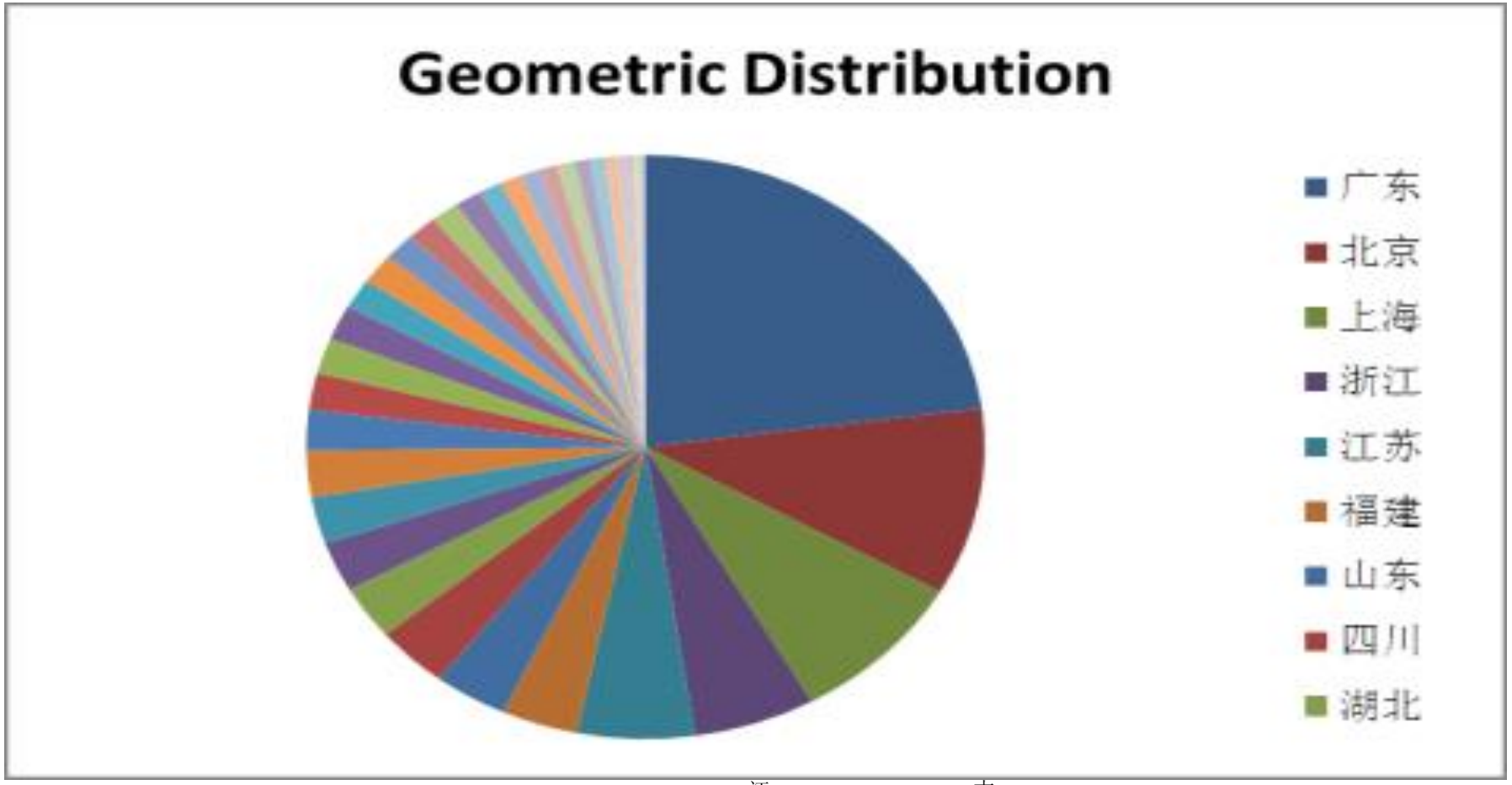
不同类型用户的分布

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

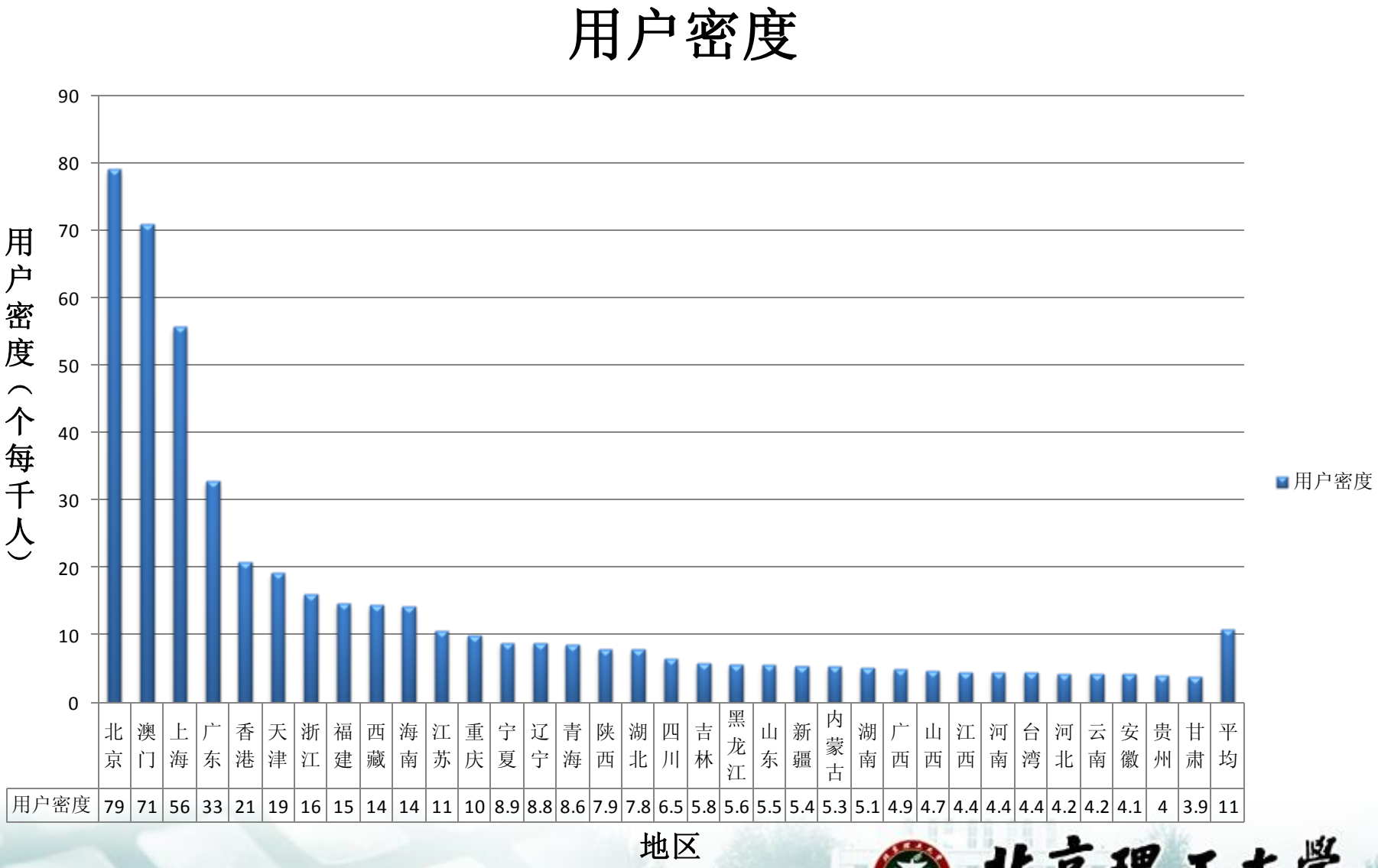
总用户数



江 古

地区

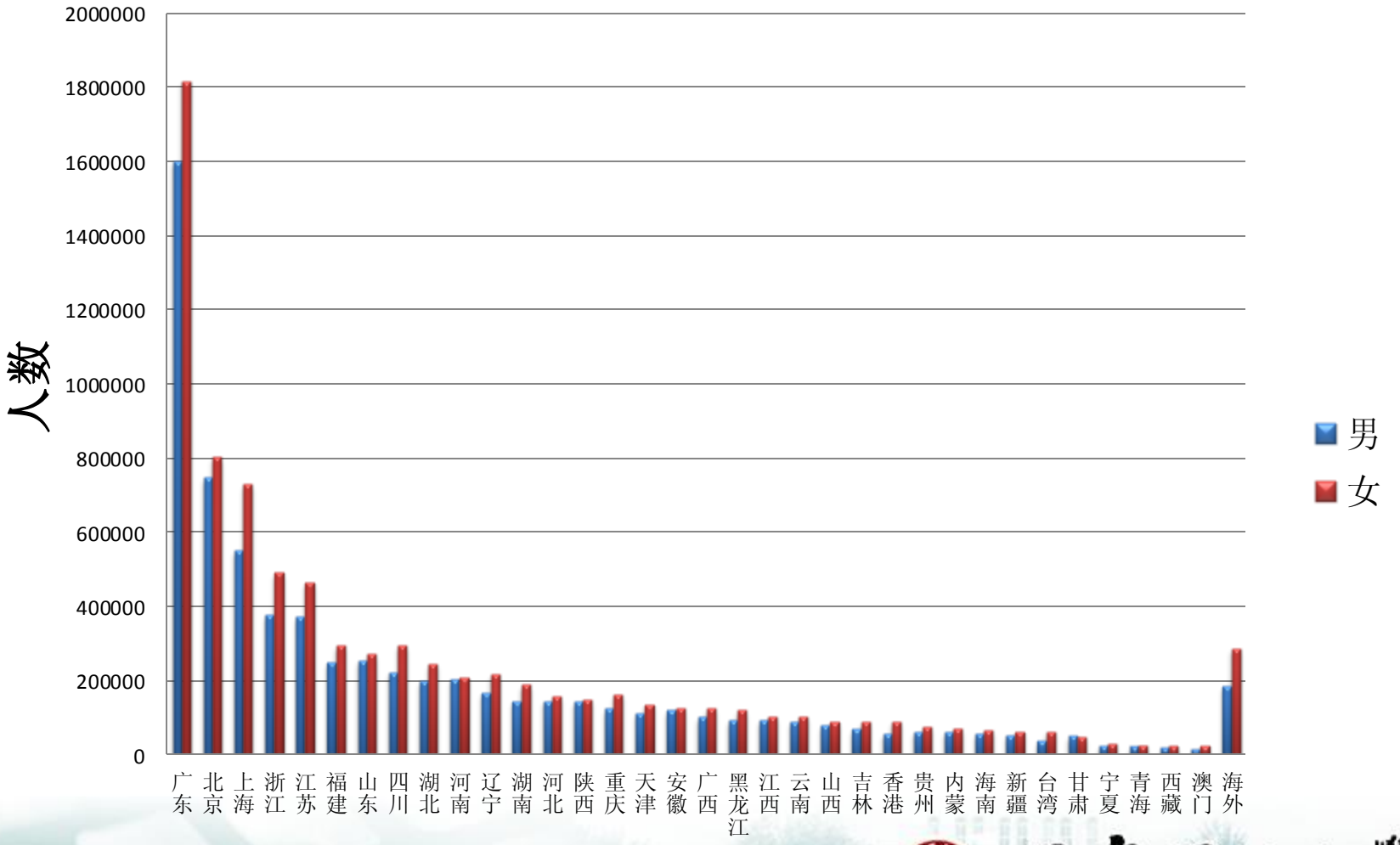
不同地区的微博用户密度



	北京	澳门	上海	广东	香港	天津	浙江	福建	西藏	海南	江苏	重庆	宁夏	辽宁	青海	陕西	湖北	四川	吉林	黑龙江	山东	新疆	内蒙古	湖南	广西	山西	江西	河南	台湾	河北	云南	安徽	贵州	甘肃	平均
用户密度	79	71	56	33	21	19	16	15	14	14	11	10	8.9	8.8	8.6	7.9	7.8	6.5	5.8	5.6	5.5	5.4	5.3	5.1	4.9	4.7	4.4	4.4	4.4	4.2	4.2	4.1	4	3.9	11

性别/区域比例联合分布

地区——性别比例联合分布



北京理工大學
BEIJING INSTITUTE OF TECHNOLOGY

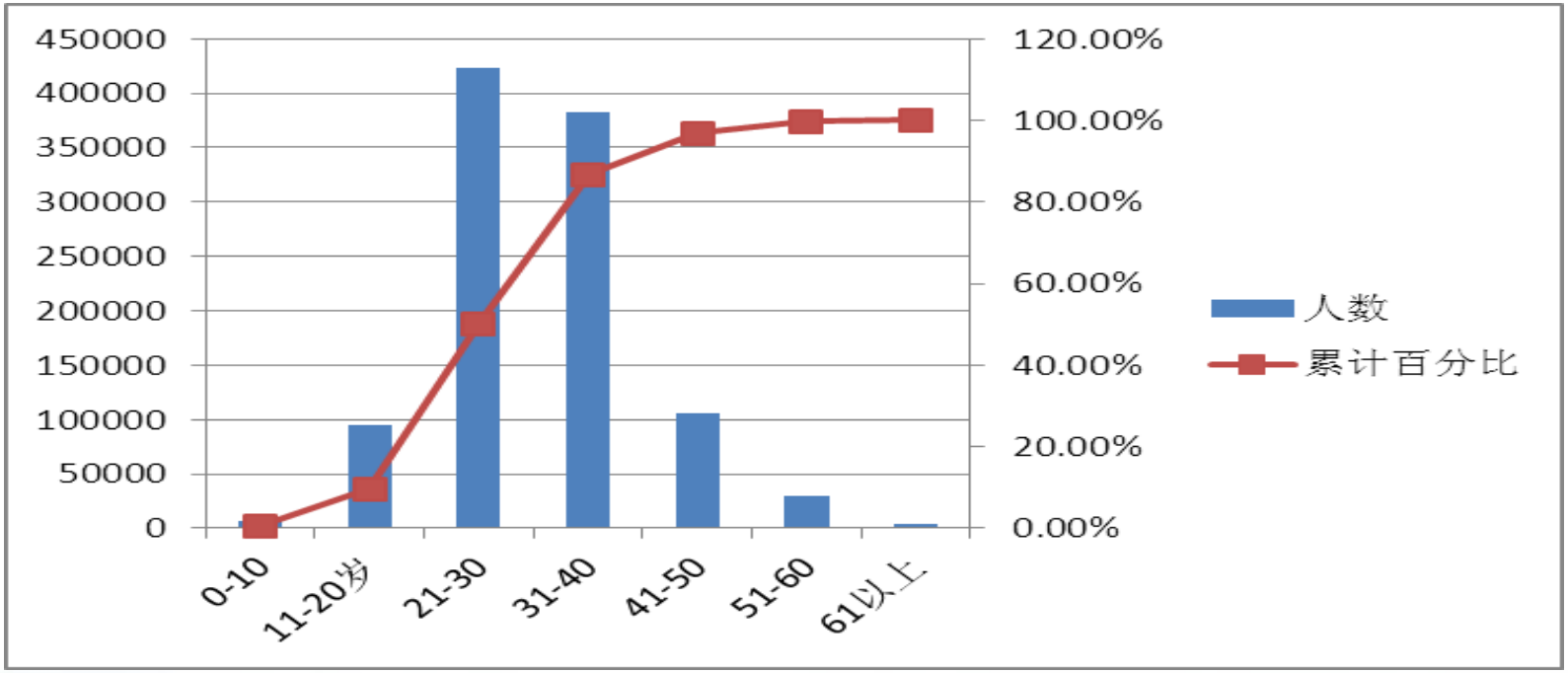
省市区划内微博用户数与GDP正相关

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

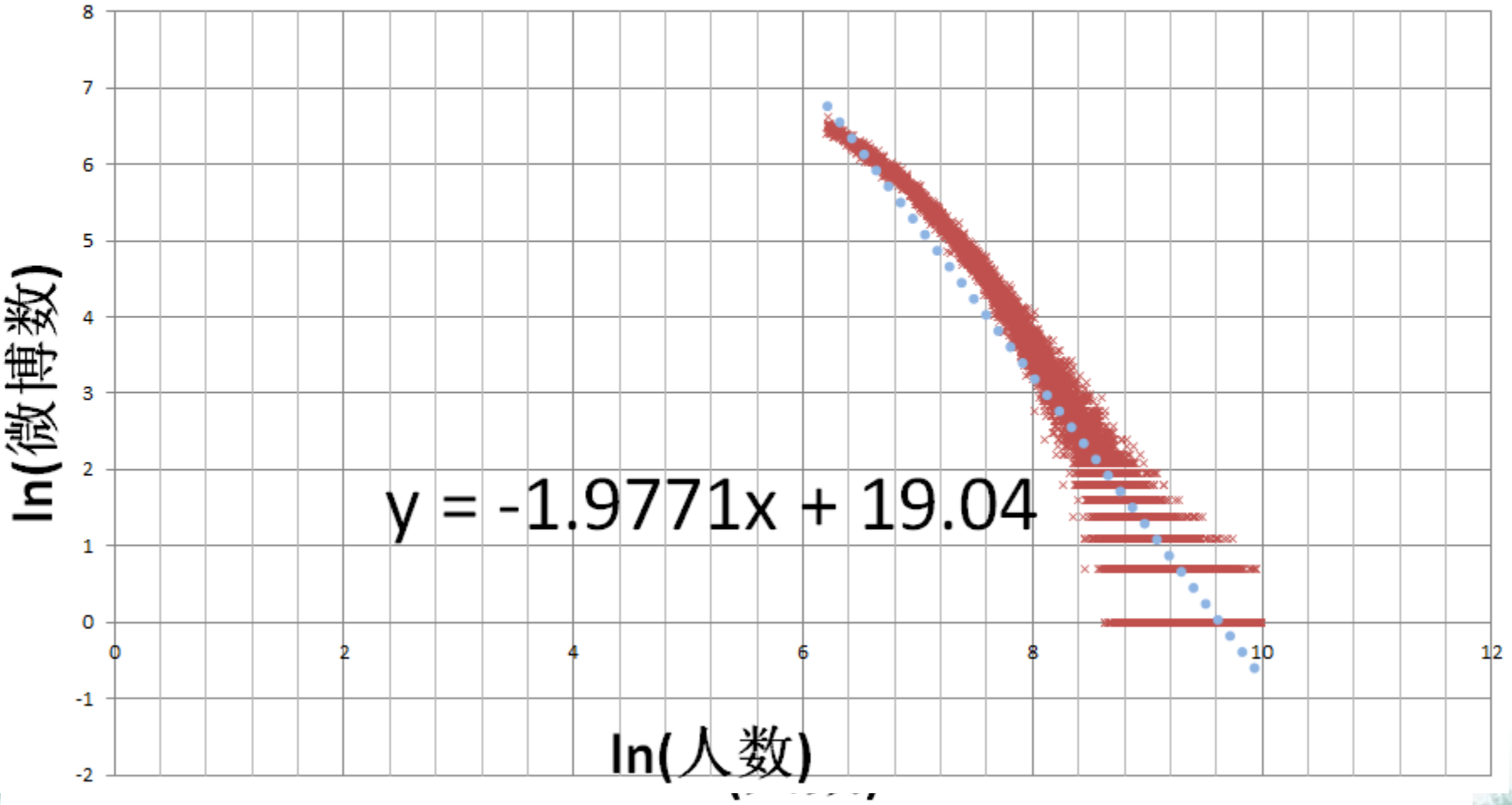
教育/年龄挖掘

- 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%。

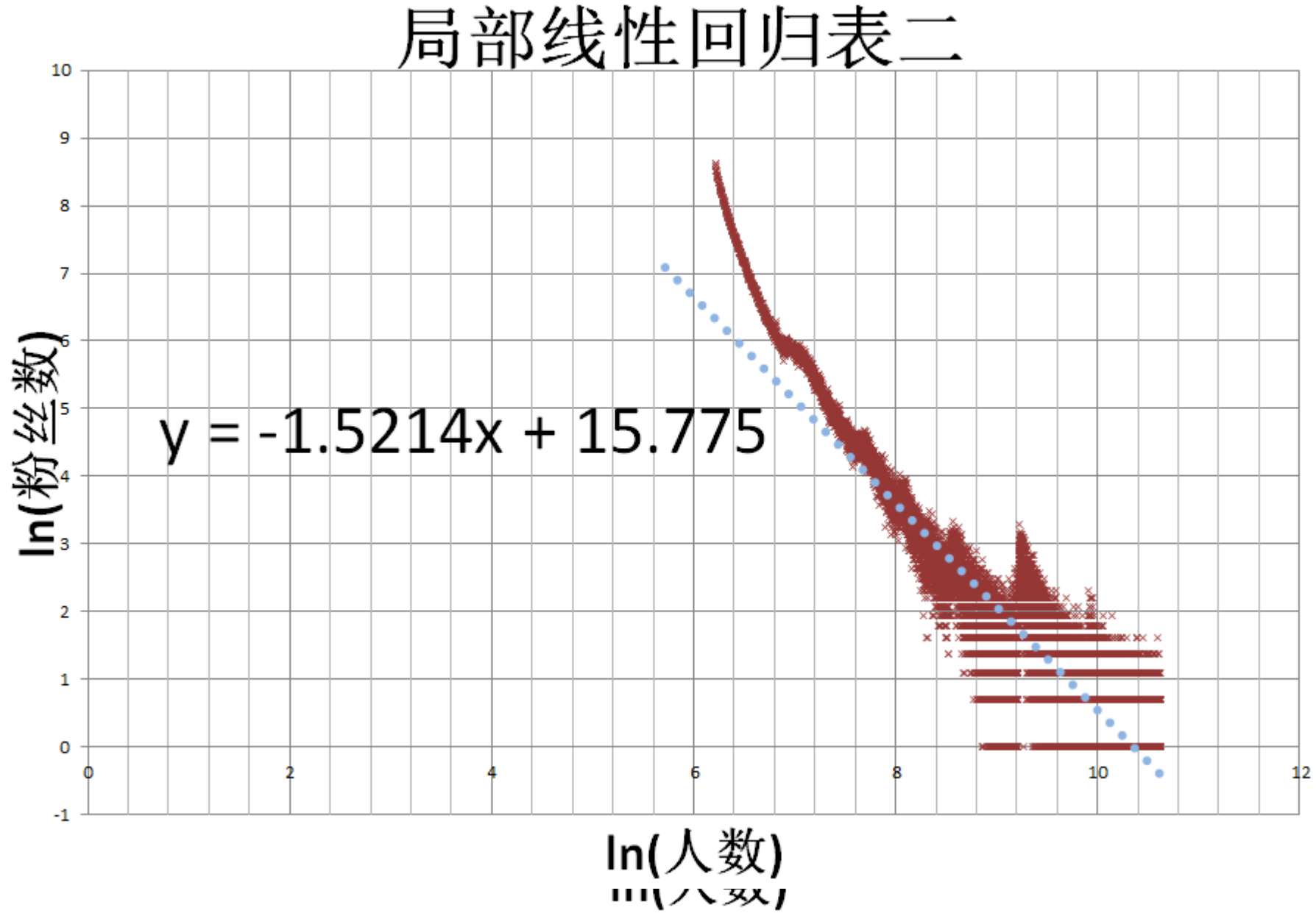


微博发布数规律

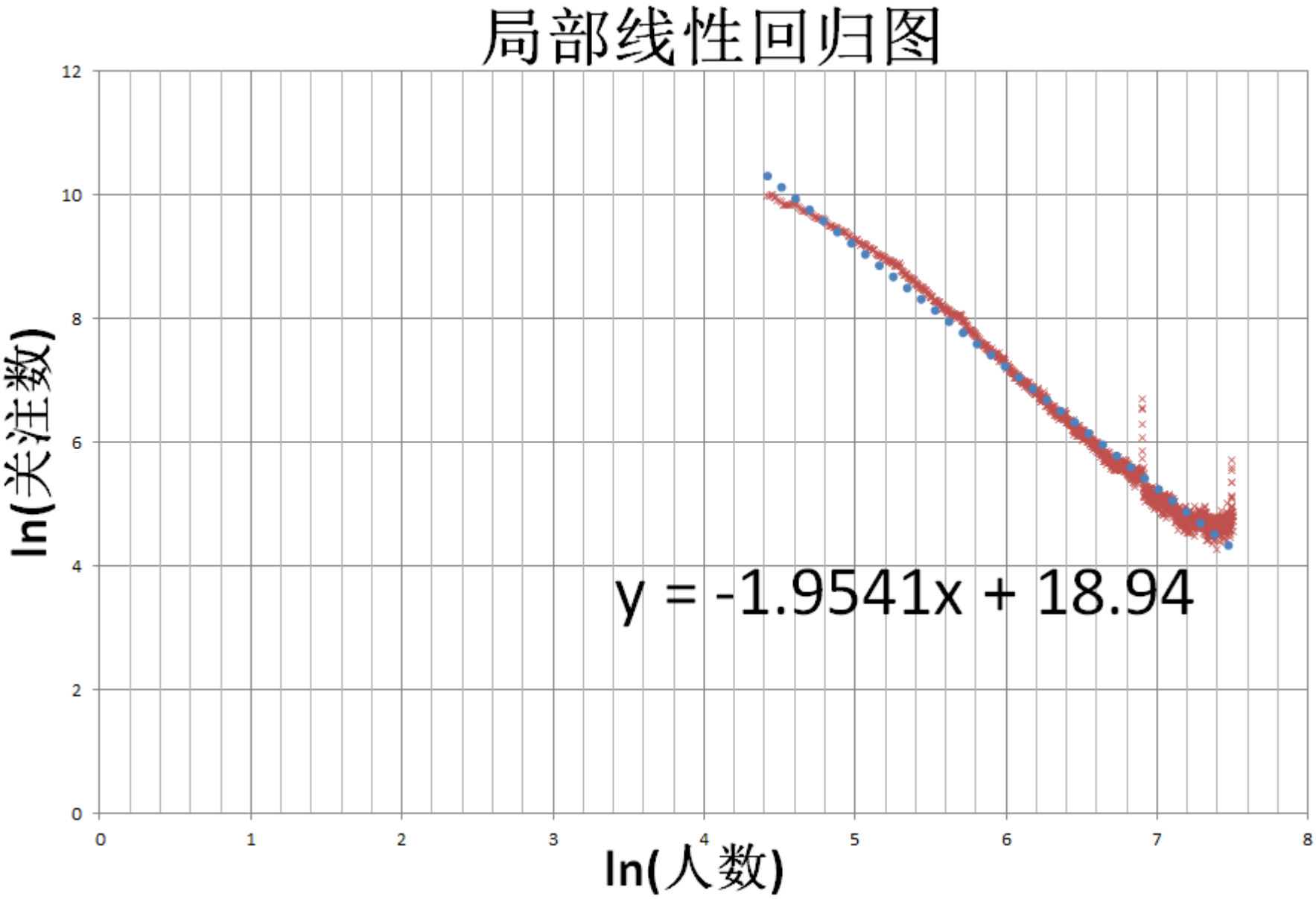
局部线性回归图二



微博粉丝数规律


局部线性回归表二
y = -1.5214x + 15.775
ln(粉丝数)
ln(人数)
10
9
8
7
6
5
4
3
2
1
0
-1
0
2
4
6
8
10
12

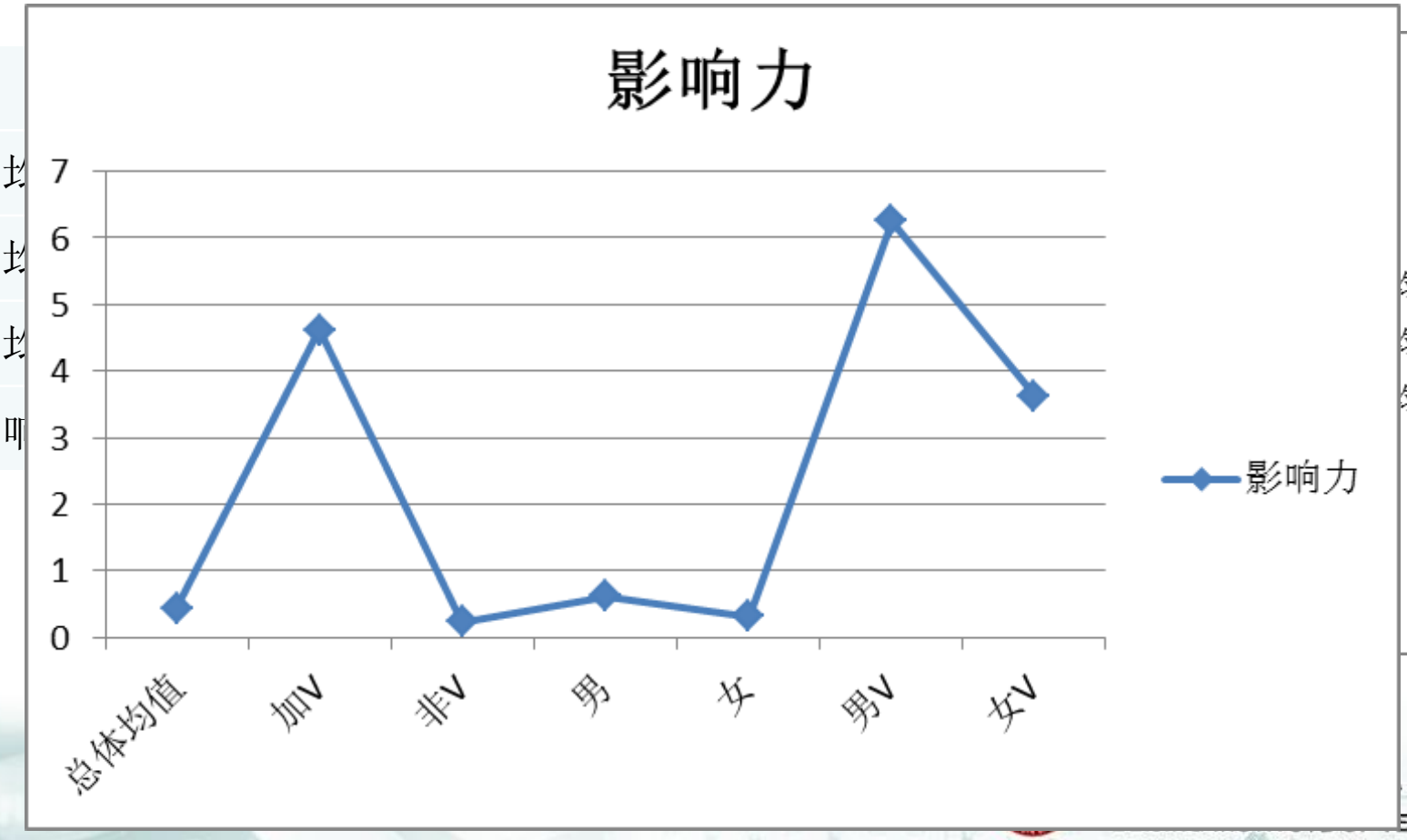
关注数规律



不同类型用户的影响力分析

影响力计算算法：

- 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。
- 已经发布新浪微博应用“微博个性热词云”：[illegible]的个性，并计算不同主体个[illegible]；并研究个体兴趣的迁移变化[illegible]
- http://esyrt.sinaapp.com/

微博个性分析

- 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
  - 输出个性化词云

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),)

博主个性化建模：沈阳教授

北京理工大學
BEIJING INSTITUTE OF TECHNOLOGY

张华平的个性化特征演化

2011年9月20日

张华平的个性化特征演化

2012年2月25日

北京理工大學
BEIJING INSTITUTE OF TECHNOLOGY

张华平的个性化特征演化

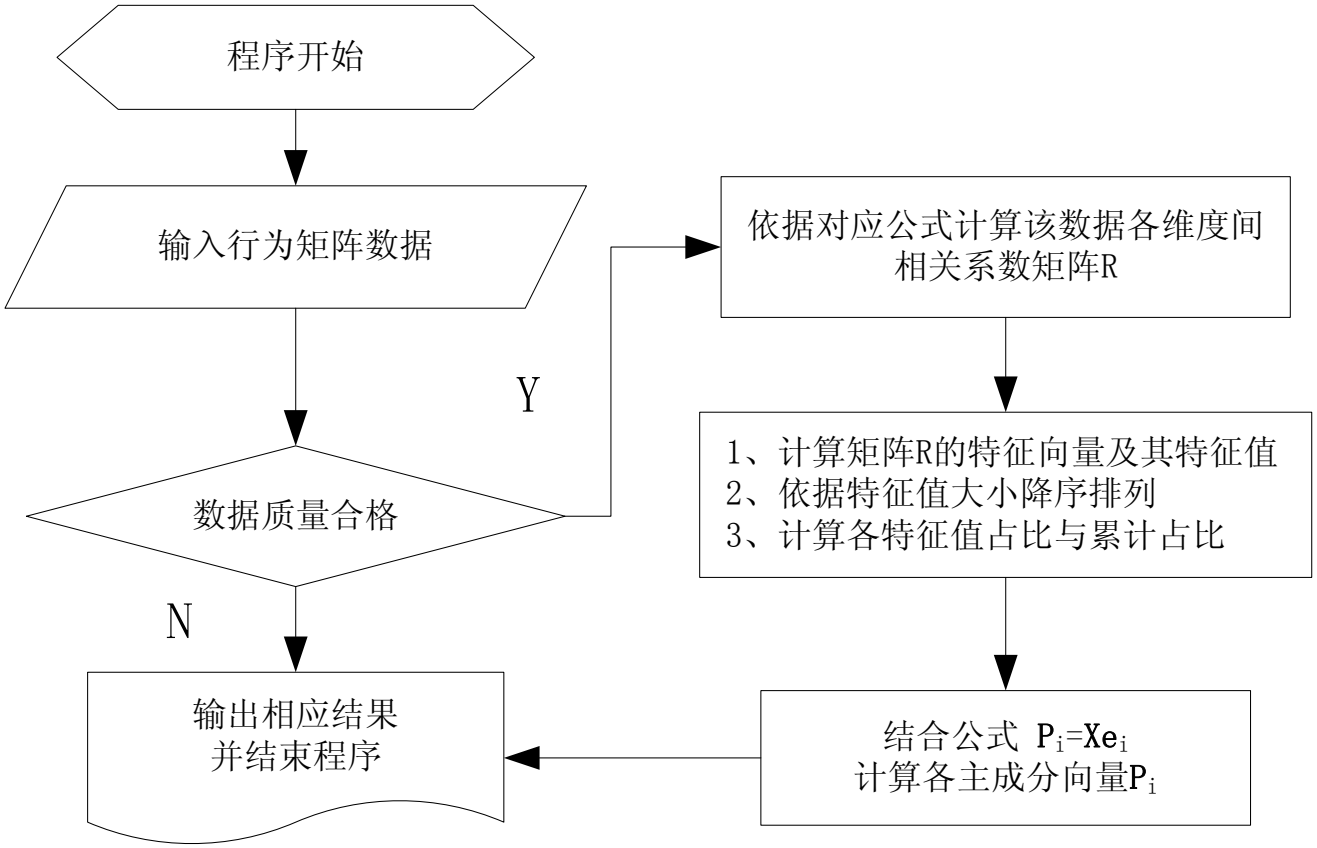
2013年4月23日

微博博主微观行为建模

焦点定位

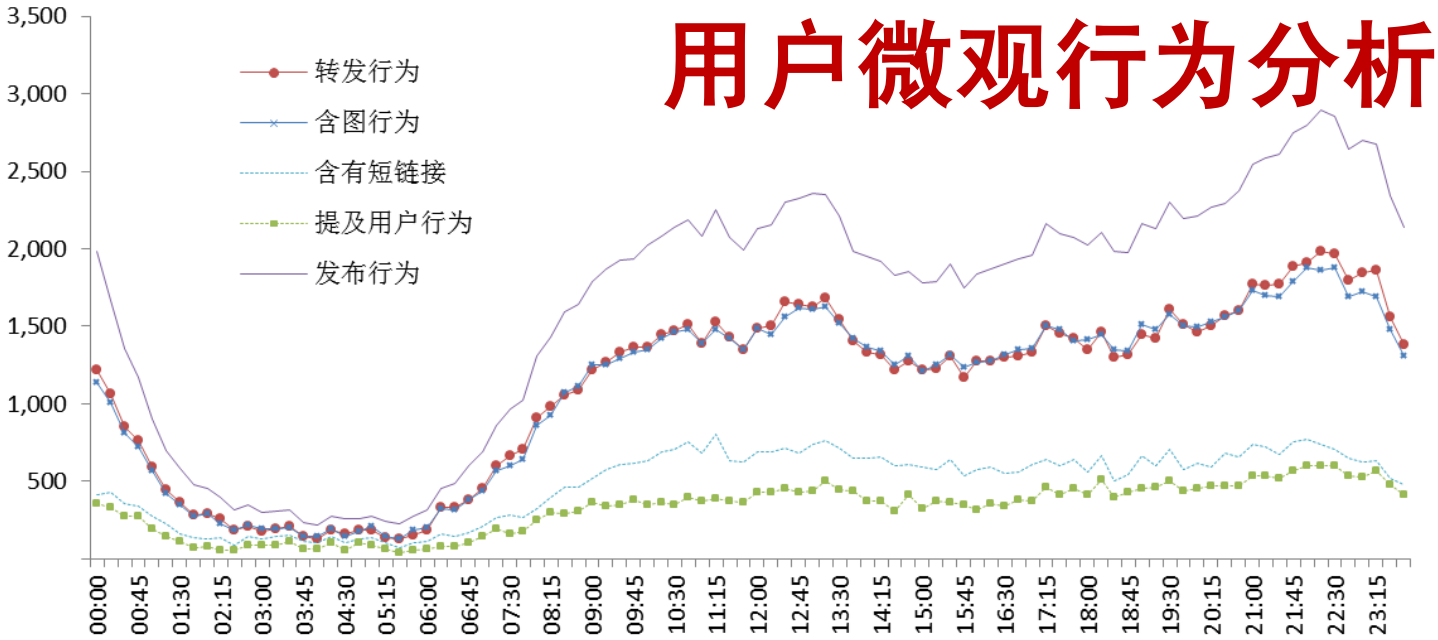
日期\时段	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	总计
2011/1/1	0	2	0	2	0	0	1	1	0	1	5	0	0	0	0	0	0	0	1	0	0	0	0	0	13
2011/1/2	0	3	4	0	0	2	0	2	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	14
2011/1/3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	3	0	1	5
2011/1/4	1	1	0	0	2	0	0	1	0	0	0	0	0	0	0	0	1	1	0	1	0	0	0	0	9
2011/1/5	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	3	2	1	0	0	0	0	16
2011/1/6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	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	0	2	1	0	1	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	0	1	3	1	2	0	1	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

行为矩阵分析系统流程

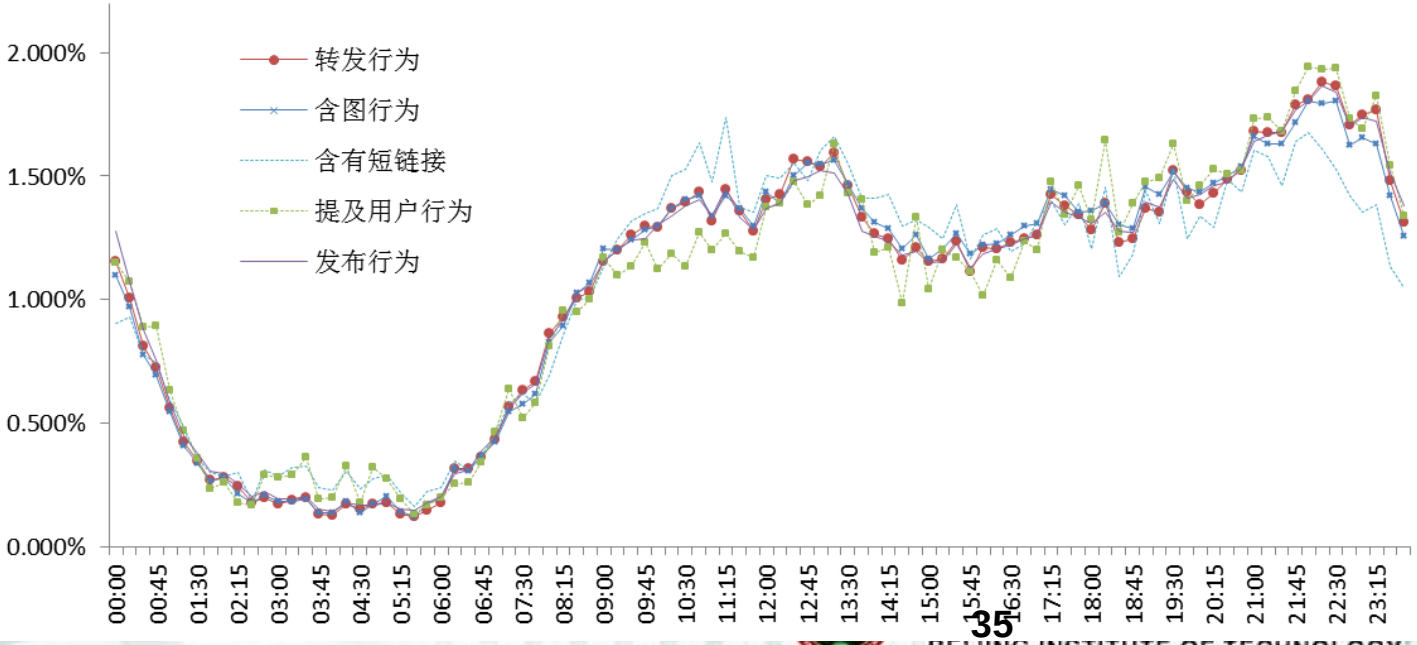


用户微观行为分析

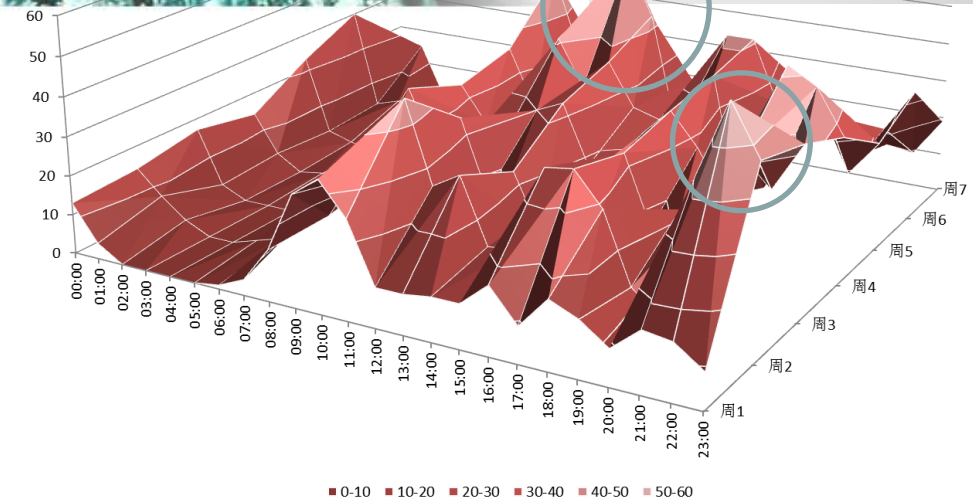
第一主成分
（原始数据）



第一主成分
（归一化）

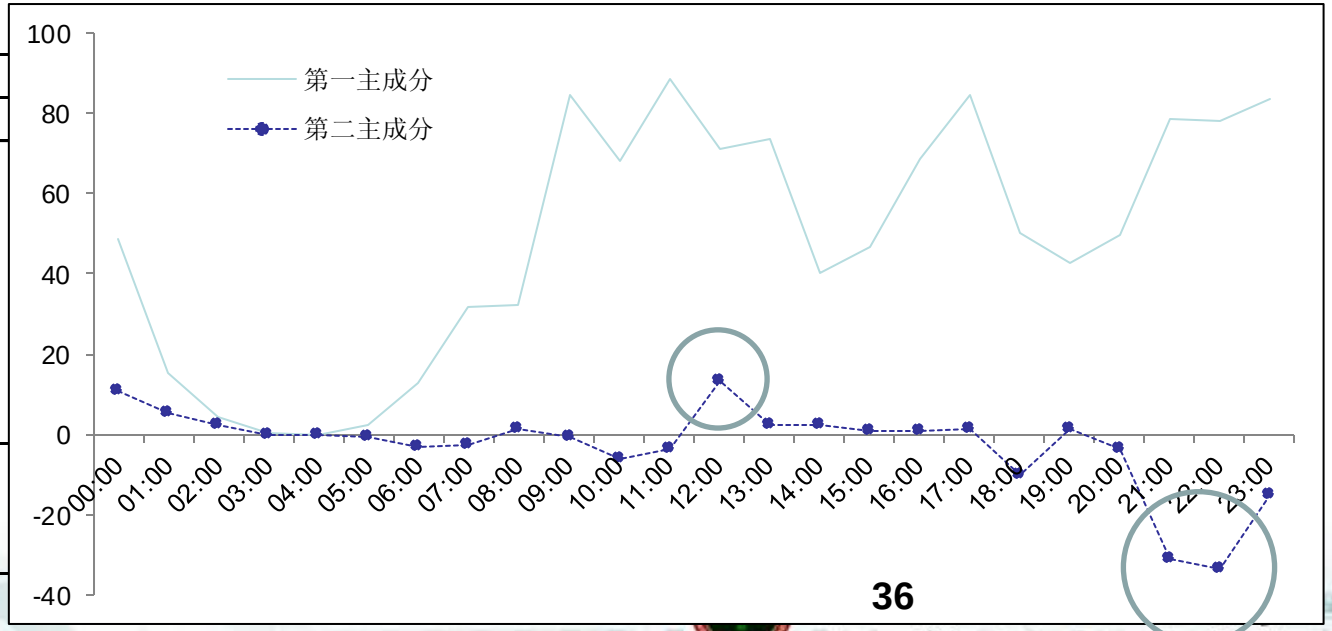


用户微观行为分析



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

	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

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

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

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



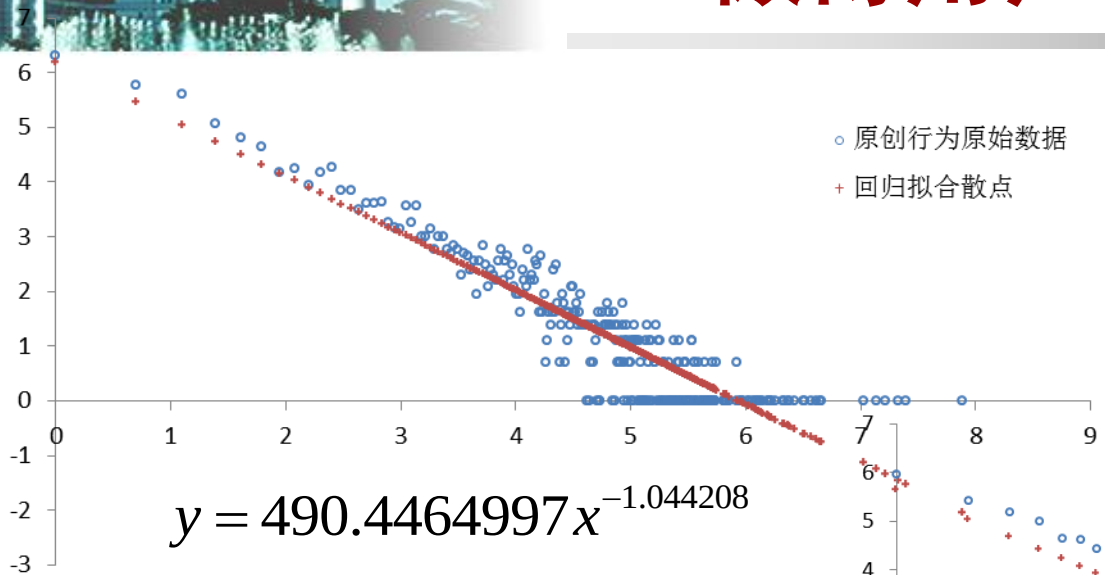




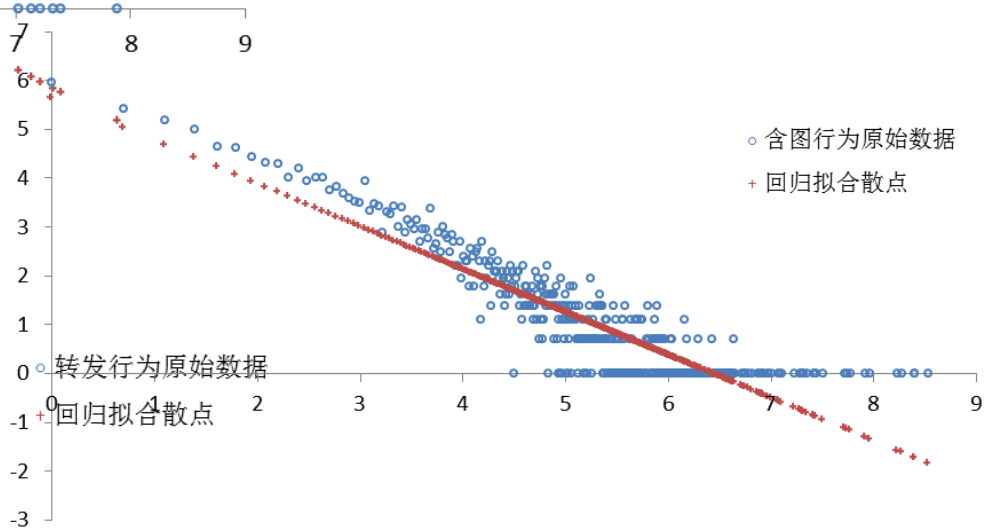

	张华平	任志强	潘石屹	张鸣	白硕	林伯强	张栋	方文山	刘强东
张华平	1								
任志强	0.447339	1							
潘石屹	0.746915	0.760761	1						
张鸣	0.84744	0.612968	0.818428	1					
白硕	0.698806	0.644066	0.81019	0.704533	1				
林伯强	0.603462	0.343865	0.602498	0.813252	0.482863	1			
张栋	0.773073	0.465831	0.758745	0.765128	0.843614	0.700826	1		
方文山	0.073967	-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

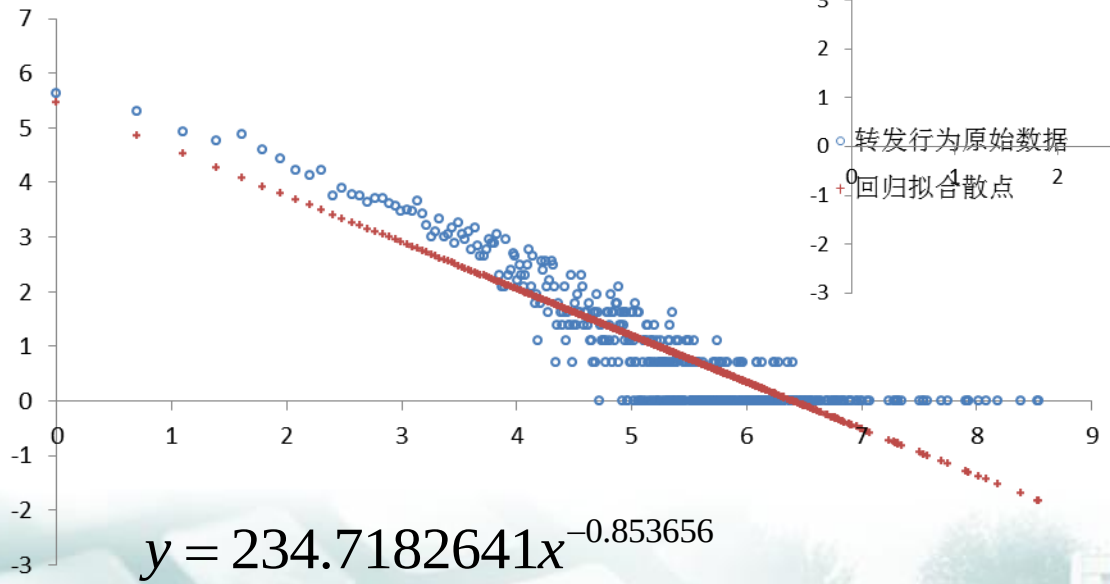
（5000人数据）



$$y = 490.4464997x^{-1.044208}$$



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



$$y = 234.7182641x^{-0.853656}$$

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：

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

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

$$\boldsymbol{\pi} = \boldsymbol{F}^T \boldsymbol{\pi}$$

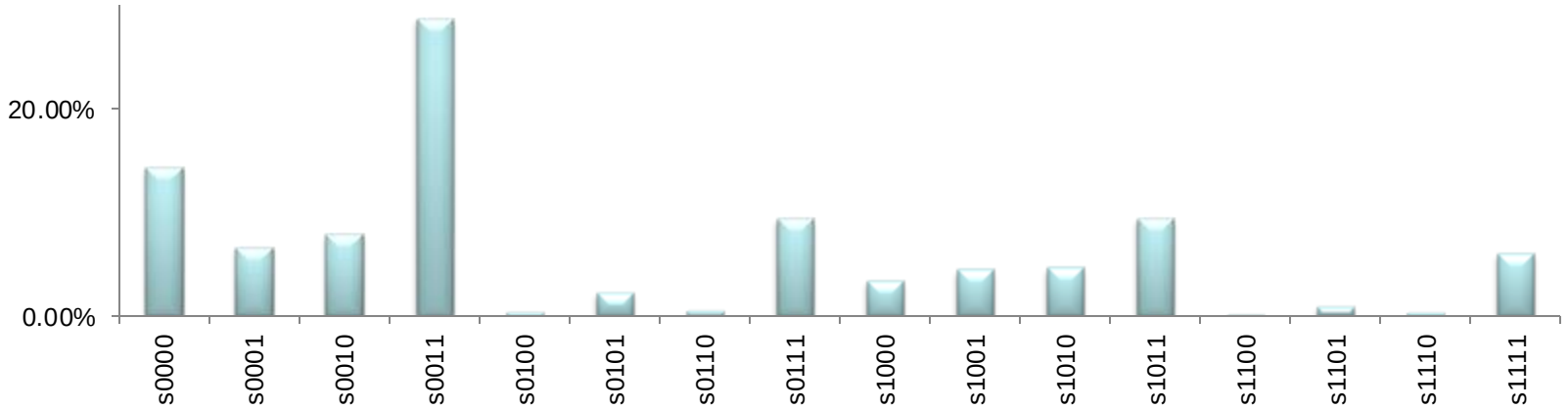
存在唯一解，也即特征值 1 所对应的特征向量。

微博用户行为转换的建模

将该向量进行归一化处理后，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



北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

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；
  - 计算关键语义
  - 综合计算词分

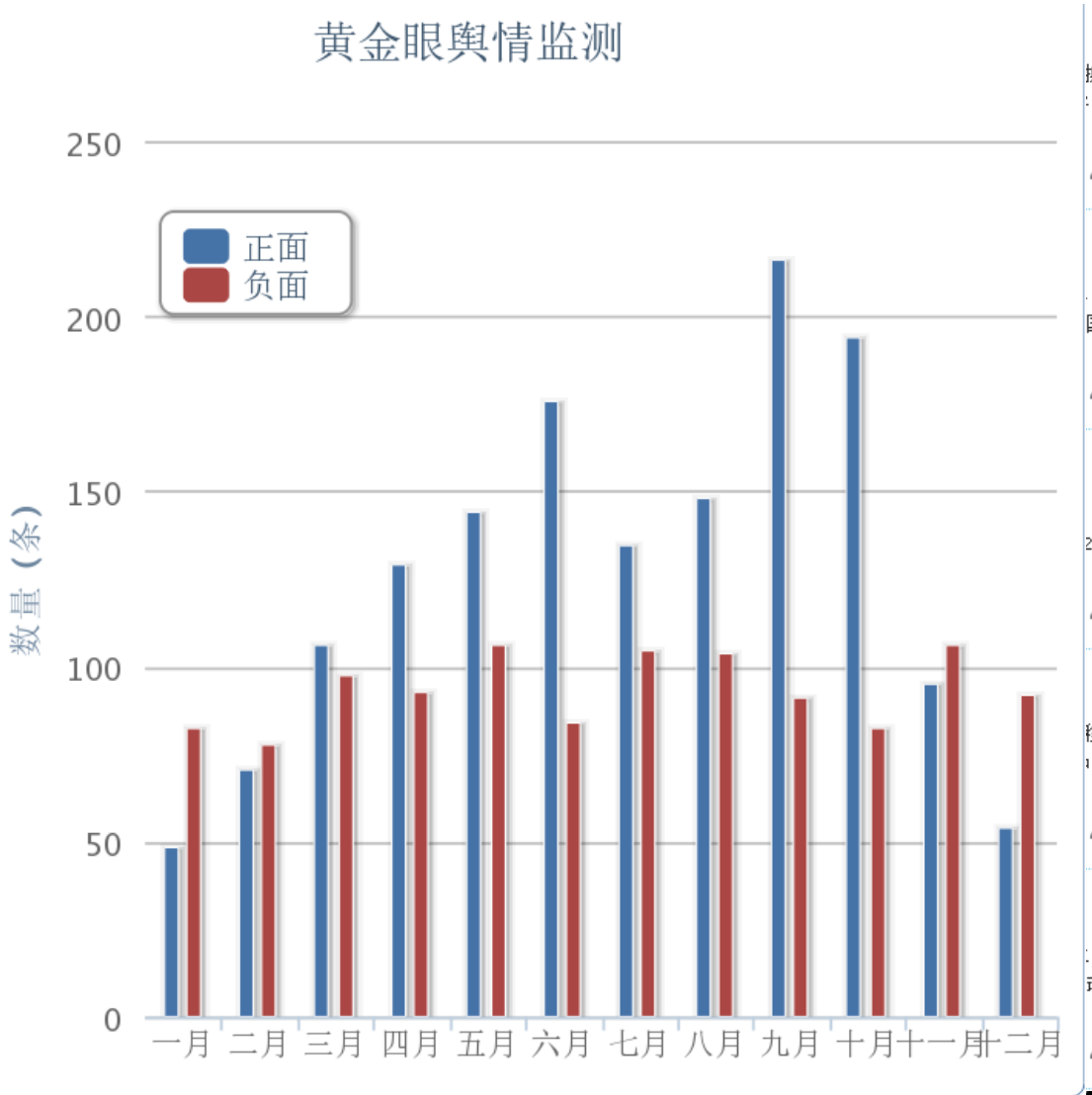
- 微博情感分析与博主情绪感知

微博话题跟踪

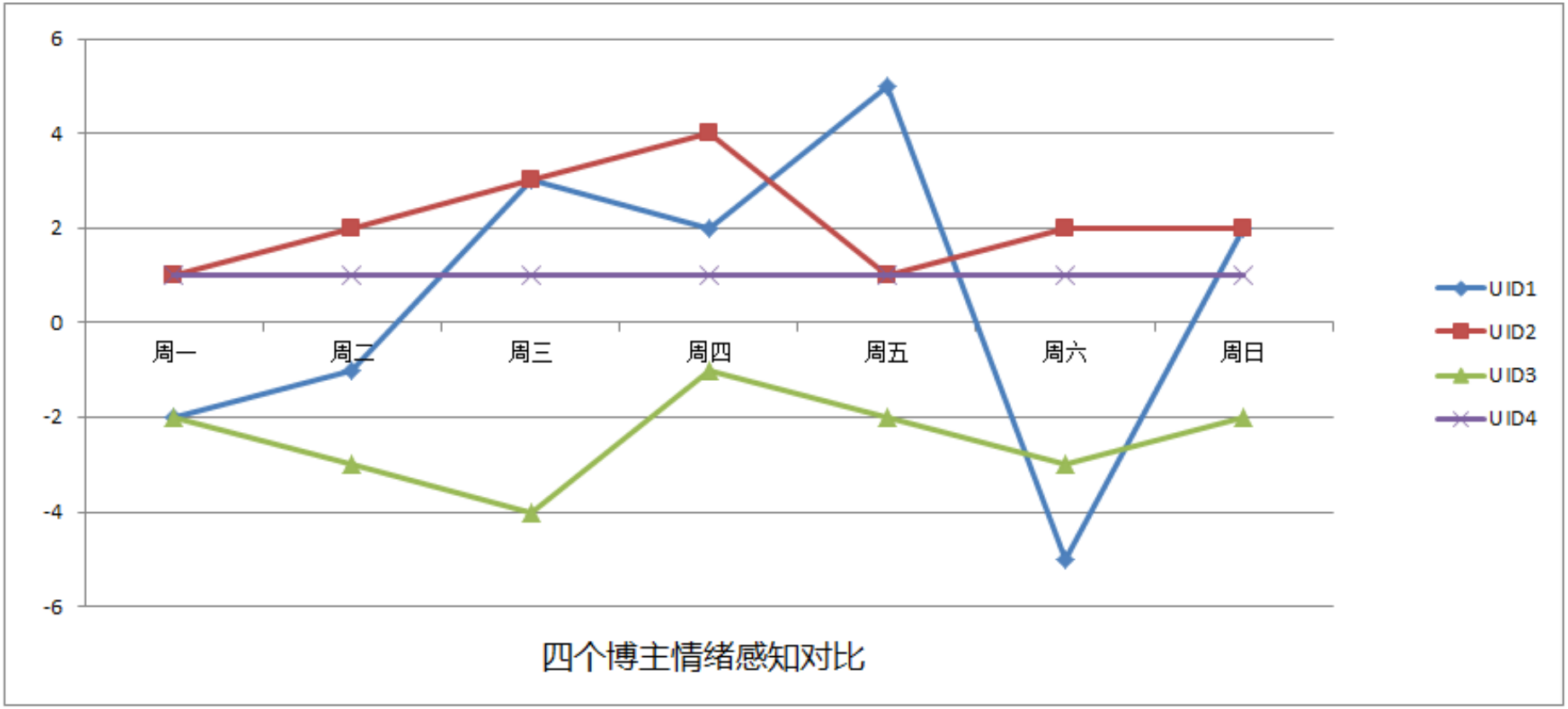

一线
车辆
也不会让
震后
全力保障
没有
他们
客服
救灾
社会
捐款
级地震
第一时间
帐篷
新浪
联通
公益
赶赴
但却
自己
公司
雅
物资
祈福
平台
改变
能改变
分钟
捐赠
联系
都是
扩散
爱心
自
专业
准备
信息
芦山
客户端
新闻
物流
人民
企业
在行动
紧急
应急
灾区
微博
安
万元
发布
天
加油
现在
官方
中国
电话
医院
灾难
雅安地震
地
CCTV
360
人员
宝兴县
余震
3G
成都
地震局
志愿者
工作
可以
四川
我们
电信
倒下
如果
服务
中心
看到
救援
龙门
帮助
截至
通信
交通
用了

“明码暗语”识别

微博情感分析


黄金眼舆情监测
正面
负面
数量（条）
250
200
150
100
50
0
一月
二月
三月
四月
五月
六月
七月
八月
九月
十月
十一月
十二月

微博博主情绪感知



总结

- 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；
- 微观：从个人的微博内容与行为矩阵，建立个性与[illegible]型；
- 内容：从内容理[illegible]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

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

北京理工大學
BEIJING INSTITUTE OF TECHNOLOGY