大数据时代的社会化新媒体舆情

Research on New Social Media in Big Data Era

张华平 博士 副教授

大数据搜索与挖掘实验室

kevinzhang@bit.edu.cn

www.nlpir.org

2018.12

Gov1.0遭遇Web3.0

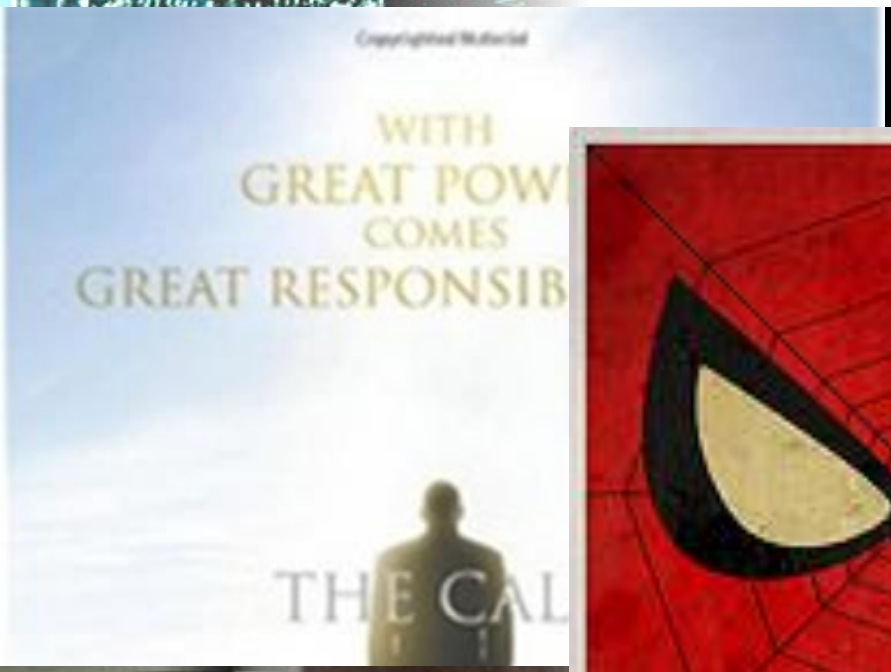
天涯论坛 bbs.tianya.cn

YouTube

新浪微博 weibo.com

twitter

传播如何入脑入心？



灾难报道之花果山地震

垮了18洞穴

死了多少猴狲?

只有500颗桃树被埋。

到底死了

活着的猴狲

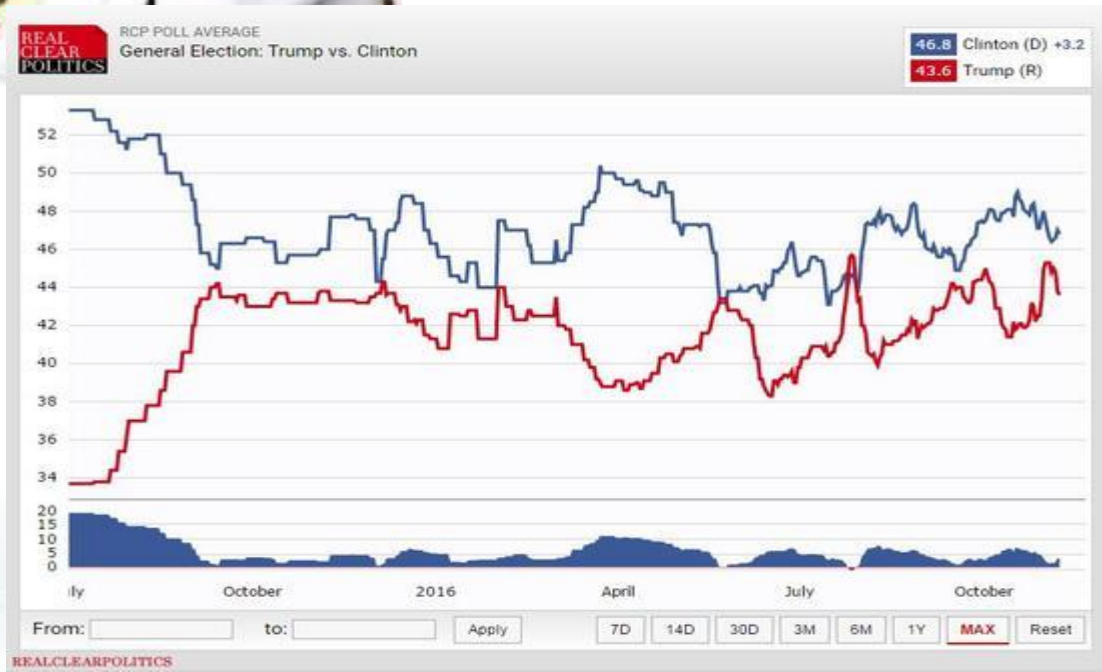
了

如实道来，到底死了多少?

猕猴情绪稳定，对灾后重建充满信心。

已成立医疗专家组，观音现场指挥，感谢领导关心支持

小数据精英 VS 大数据庶民



小数据精英 vs 大数据庶民

从棱镜手机监控看大数据洞察力...

纲要

大数据
新媒体

I 新媒体传播创新

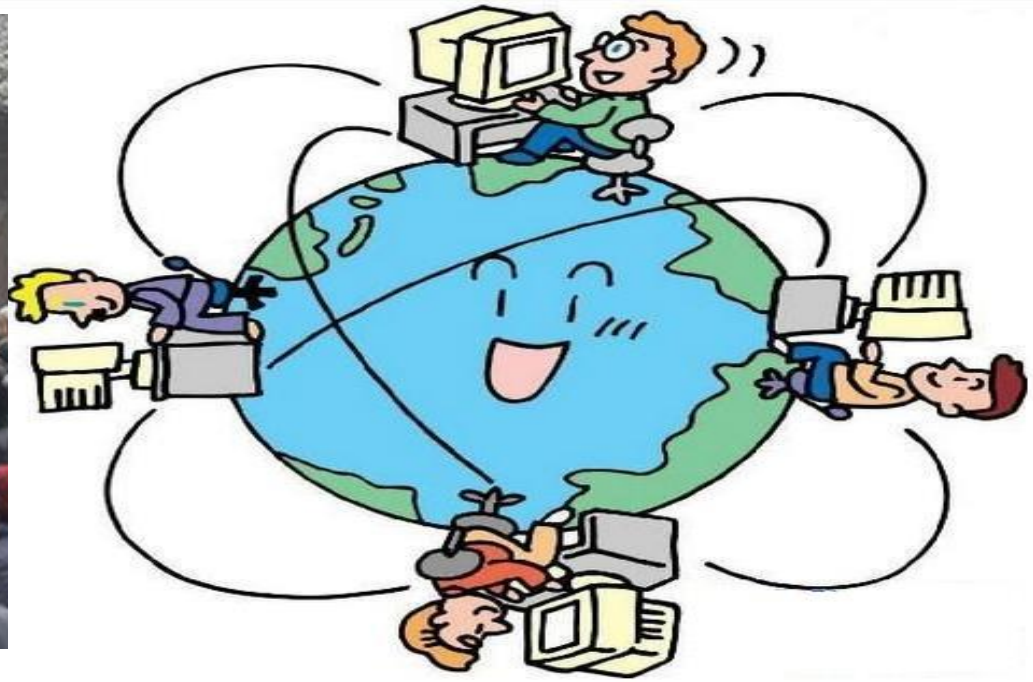
II 社会化新媒体舆情特点

III 社交媒体分析关键技术

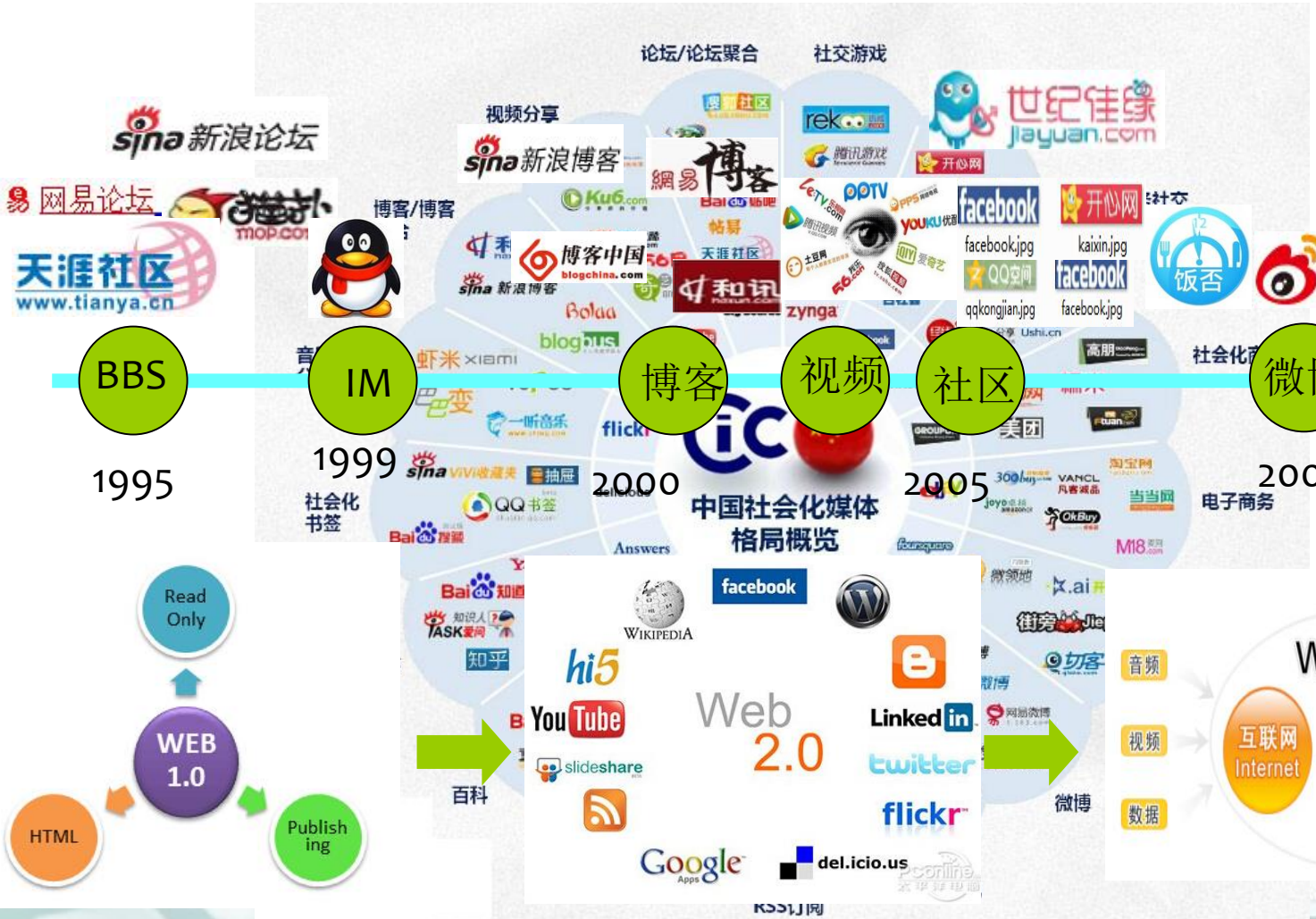
IV 大数据新媒体应用实战案例

社会化媒体

- 社会化媒体（社交媒体）运用易涉入和传播的沟通技术并以社会化交流为目的的媒体。特点：社会关系+传媒



社会化媒体发展历程


论坛/论坛聚合
社交游戏
视频分享
博客/博客
社会化书签
电子商务
中国社会化媒体格局概览
BBS
IM
博客
视频
社区
微博
微信
1995
1999
2000
2005
2009
2011

Read Only
WEB 1.0
HTML
Publishing

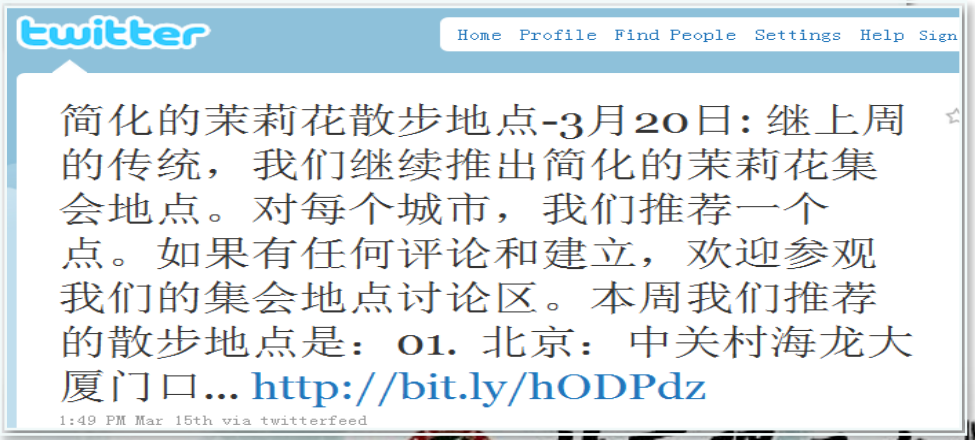
facebook
WIKIPEDIA
hi5
YouTube
Web 2.0
LinkedIn
twitter
slideshare
flickr
Google
del.icio.us
百科
微博


WEB 3.0
音频
视频
数据
互联网 Internet
通讯网络 Instant Communications
固定电话
移动电话
IPTV
PC

新媒体在社会管理中兴风作浪

利比亚Facebook革命

Twitter煽动 “茉莉花革命”



传统媒体 vs. 新媒体

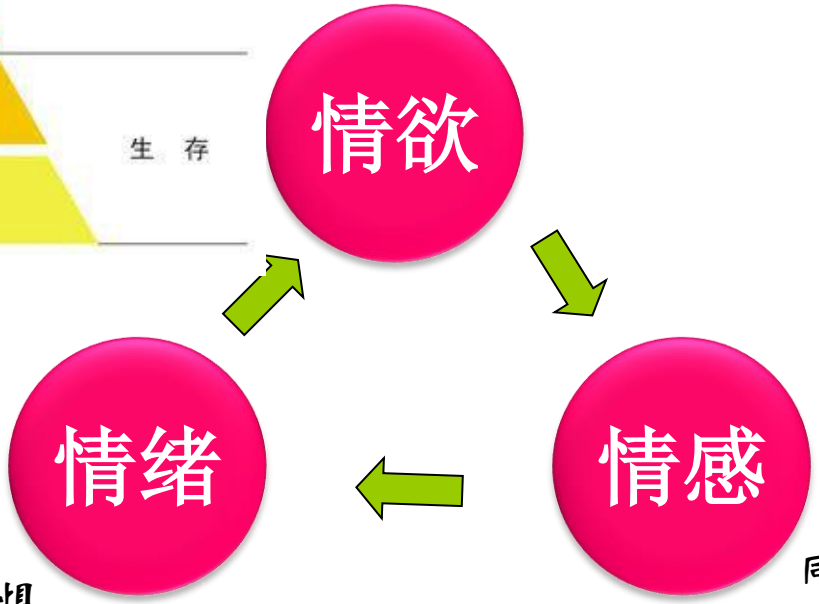
传媒时代	报纸/电视	互联网1.0	新媒体
内容	正式	半正式	非正式
传播方式	一对多广播，无反馈的；	少对多浏览，弱化社交	多对多，社交型，
主体	授权机构，少数	大部分网民	几乎所有人
受众	被动接受，参与感弱	主动获取，部分参与	主动推送，收发全参与
生产过程	先审后发	先发后审	即发少审
时机/速度	24-72小时	1-2小时	即时，快且影响面广
代表	人民日报，CCTV	新浪新闻，博客，	微博,微信，facebook
场景	政府宣传，传教	小范围演讲互动	对等交流

社会媒体传播实战技巧：媒介

舆论的传播途径

社会媒体传播实战技巧：内容

马斯洛需求层次理论



艾克曼：喜、怒、哀、惧
仇官仇富情绪

同情与代入感

社会媒体传播实战技巧：内容

1. 代入感原则，拒绝自嗨

社会媒体传播实战技巧：内容

2.拉家常讲故事，拒绝高大全空洞说教

收藏 | 转发 138 | 评论 112 | 赞 451

熬夜看球

社会媒体传播实战技巧：内容

3. 角色个性化拒绝平庸：幽默风趣，借题发挥

0
关注

Lv.30

其他

大数据

对于闻惯了城市气

7月11日 08:47 来自

收藏

“有干部问，既然我
要有几个反面典型
今天的幸福。这就

7月10日 21:34 来自

从此，朝鲜人民的幸福又多了一条，没有股灾。

@作家崔成浩

炒股的赶紧把电脑屏幕倒过来吧！

7月8日 09:37 来自 Koryolink iPhone 6 Plus　转发 512　评论 579　1306

7月8日 09:47 来自 Koryolink iPhone 6 Plus

收藏　转发 212　评论 312　787

炒股的赶紧把电脑屏幕倒过来吧！

7月8日 09:37 来自 Koryolink iPhone 6 Plus

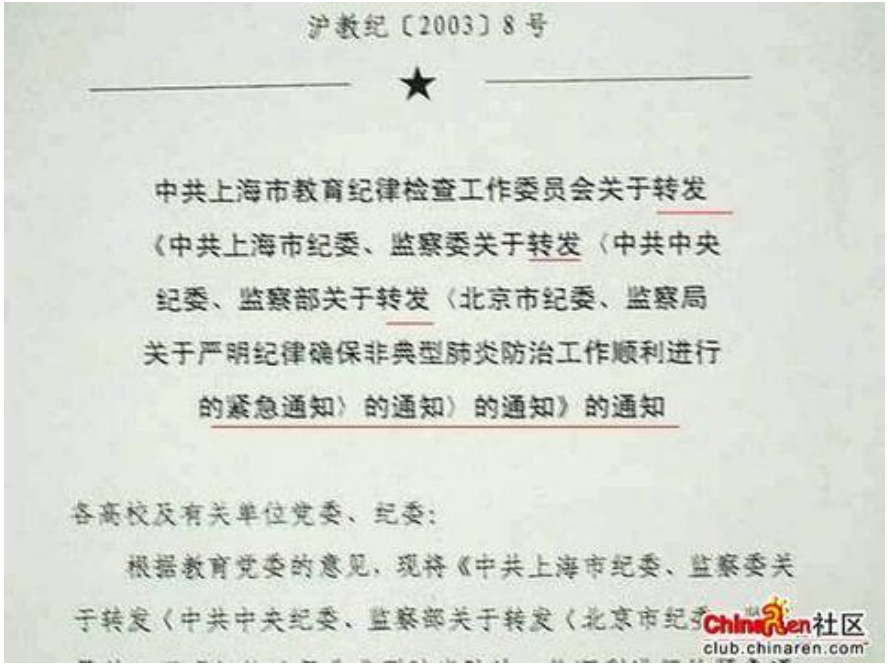
收藏　转发 512　评论 579　1306

请问，长生不老药的成份是什么？

社会媒体传播实战技巧：内容

- 4.主题创意（幽默、借用，故事性）、好记易传播（3-4音节）

38元大虾
微笑局长
表叔
房姐
我爸是李刚
土豪，我们交朋友吧
光盘计划



沪教纪〔2003〕8号

★

中共上海市教育纪律检查工作委员会关于转发《中共上海市纪委、监察委关于转发〈中共中央纪委、监察部关于转发〈北京市纪委、监察局关于严明纪律确保非典型肺炎防治工作顺利进行的紧急通知〉的通知〉的通知》的通知

各高校及有关单位党委、纪委：

根据教育党委的意见，现将《中共上海市纪委、监察委关于转发〈中共中央纪委、监察部关于转发〈北京市纪

纲要

大数据
新媒体

I 新媒体传播创新

II 社会化新媒体舆情特点

III 社交媒体分析关键技术

IV 大数据新媒体应用实战案例

舆情生命周期剖析

impact

即时分散、主体性强、难监测而可疏导

信息与群体聚集，主体隐蔽，易监测难疏导

适于舆情事后评估，丧失了监测疏导时机

时间淡忘消解，但相关事件可再次点燃；如PX;扶不扶?

time

舆情发生期

始作俑者：当事人

舆情传播期

网络受众：舆情大V/网络粉丝(愤青)

舆情倒逼期

社会群体：公权/社会大众

舆情消化期

未来族群：社会心理（妖魔化）

舆情分析困局迷思

以“舆情”为中心的舆情分析与监控困局

- 四处扑火，防不胜防；
- 全面监控不留死角技术不可行，经济上不现实；
- 所谓舆情系统或者舆情监管仅仅实现了舆情事后的监测分析，不过是“亡羊补牢”；舆情呈现泡沫化倾向；
- 舆情千变万化，转世党层出不穷，关键词变种让人脑洞打开；预测预警几乎不可能
- 部分舆情分析以“威胁讹诈”为主要手段，铲事为主要诉求；

传统舆情为中心的分析渐入困境！

NLP：自然语言处理？身心语言程序学

造谣、软文、水军、影射、反讽

舆情情感分析、评分还能信吗？

每天用点心理学-湖南NLP学院：你要相信”当下你的选择，一定是你能做出的最好选择“我们做的任何事情，都是为了满足自己的一些需要。在那些特定的环境里，也许你事后会后悔自己当时的选择，但其实当给多你一次机会重头来过，你还是会做同样选择，因为那是你在当时的最好。想让自己学会好的选择吗？NLP可以告诉你

5分钟前 来自 皮皮时光机 转发 | 收藏 | 评论

中微子u：//@李方涛2011: 之前有NLP的ACM Fellow吗

@刘知远THU：今年ACM Fellow揭晓。http://t.cn/SqgiC0 其中Dan Roth (UIUC)和Amit Singhal (Google)是与NLP和IR相关的，关注。

12月9日01:04 来自 新浪微博 转发(5) | 评论(2)

20分钟前 来自 微博搜索 转发 | 收藏 | 评论

信诚人寿-冯艳：当我以为最年轻的NLP执行师在我们班时(18岁),花美女说她们班上有个16岁的。。。。嗯!这么早接触NLP真好!

自媒体创新的重要机遇期

- 社交网络的即时性，受众比以往任何时候更需要信息，更需要媒体；
- 自媒体众声喧哗中，受众更需要政府等权威的声音；
- 信息的廉价复制转发，谣言四起，五味杂陈中，受众更需要深度思考理性声音。

社会媒体危机公关实战技巧

- 内容：客观事实佐证，无声胜有声：
  - 表述尽可能的客观真实，不能急于撇清责任；
- 方式：以Web3.0的方式沟通：
  - 语言、方式、真诚的态度，公众容易接受的方式方法；
  - 防被炒作：慎用官方方言人，谨慎发布官方消息，不发则已，一击而中。
- 主体：合适的表述主体，一句顶一万句；
  - 社会管理机关领导、当事人、当事人直管领导
  - 第三方：社会管理机构之外的当事人、专家
  - 应当是裁判员的时候，千万别被人绑架当运动员来观摩。

社会媒体危机公关实战技巧

- 顺势：
  - 不要逆水行舟，王石汶川捐款 vs. 郭美美事件
  - 借用关注的势能，推动公众的理解、扩大美誉度。
- 时机：巧用新闻传播生命周期
  - 时间是把杀猪刀，注意力有限，古今多少事，俱付笑谈中；
  - 禁忌：逆向操作，火上浇油。
- 技术：采用较强的大数据信息技术手段辅佐
  - 对辖区对象社交网络的实时采集，精准分析，广泛获取，重点监测；
  - 舆情的搜索；
  - 热点分析、敏感点预警；
  - 用数据指数辅助研判局势；

社会媒体传播实战技巧：时机

impact

即时分散、主体性强、难监测而可疏导

信息与群体聚集，主体隐蔽，易监测难疏导

适于舆情事后评估，丧失了监测疏导时机

时间淡忘消解，但相关事件可再次点燃；如PX;扶不扶？

time

舆情发生期
始作俑者：当事人

舆情传播期
网络受众：舆情大V/网络粉丝(愤青)

舆情倒逼期
社会群体：公权/社会大众

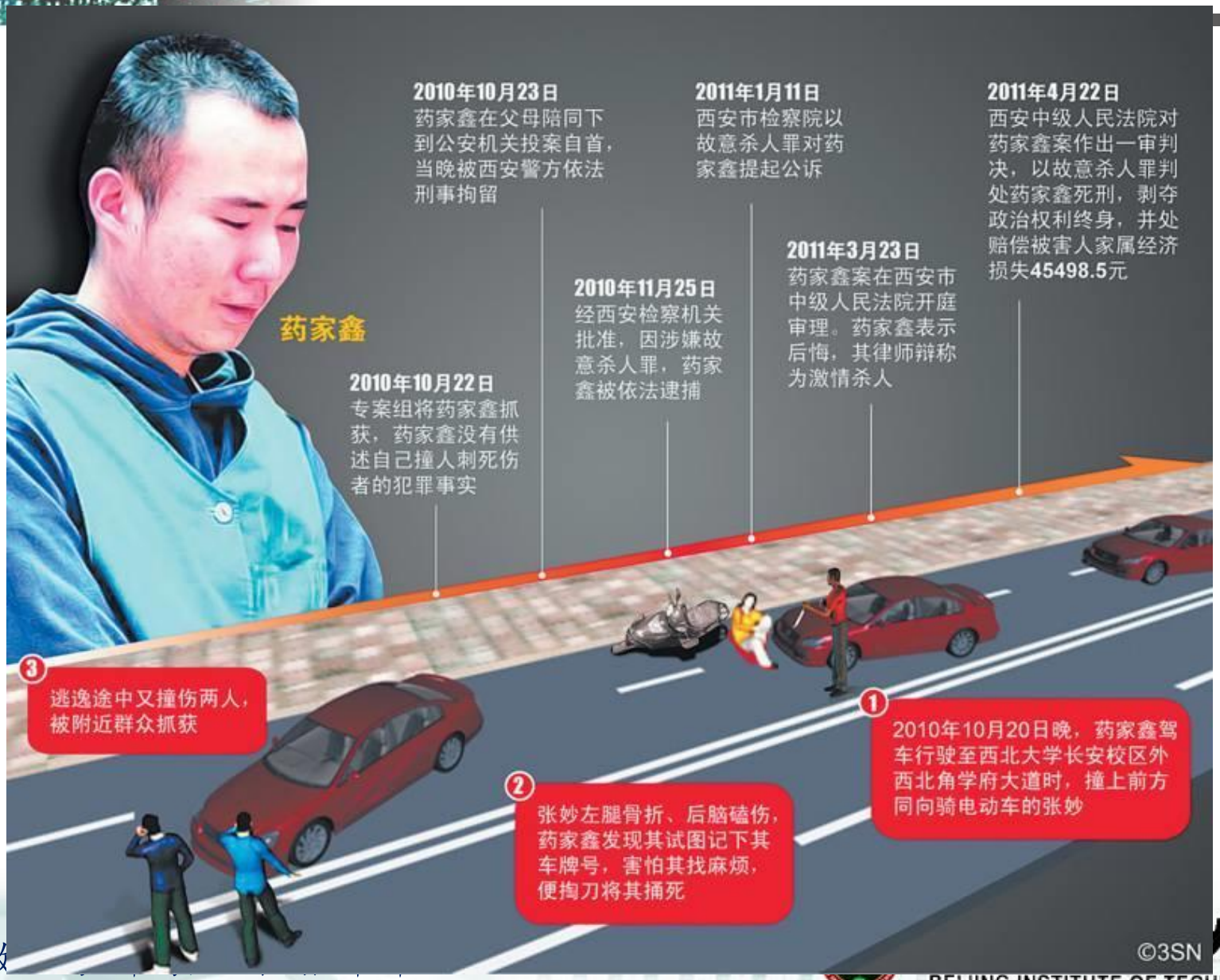
舆情消化期
未来族群：社会心理（妖魔化）

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

哈尔滨宝马撞人案

- 2003年10月16日10时事发；11月20日审理时才被媒体披露，由沈阳今报报道；没有足够的回应，对国家公平正义极大的破坏。
- **“宝马”撞人案网上点击率位居第一超过非典**
- 不是嘘的一声，而是轰的一声；不是意见领袖振臂高呼，而是陌生人成群结队

药家鑫：一出没有胜者的悲剧



药家鑫：一出没有胜者的悲剧

- 西安音 深夜，驾车
- 2010年
- 2011年 公诉。
- 2011年 广受网络调侃
- 药家鑫 百余名群众 询表。人人网 “换我，我也
- 3月23日晚，李玫瑾在央视点评称，李的说法被网友称为“钢琴强迫杀人法”，是在为药开脱罪行。李本人亦陷入漫天口水之中。

药家鑫：一出没有胜者的悲剧

鑫诉说
义母亲
起舆论
过央视
九会。
发博上
顶。

，央

国人
并抨

药家鑫：一出没有胜者的悲剧

主体	期望的结局	结局悲剧分析
药家鑫	存活保命	年仅22岁大三学生，曾经的乖乖生，因自己的犯罪丢名；同时身背骂名，成为2011年热点悲剧事件的主角
张妙	生存	本没有什么危险，但药家鑫没有给她生的机会；丢下年迈的父母，年幼的孩子
药家鑫之父	动用各种资源进行危机公关，能救子一命	老年丧子，孩子命丧黄泉，尸首都无法见到，受害者不能原谅，后期的忏悔痛心得到部分人的谅解，但是仍然广受争议
张妙丈夫父亲孩子等家属	得到公正的对待，得到药家的道歉以及相应的赔偿	张妙人死不能复生，而药家的道歉迟到了7个月，民事赔偿忽略不计，赢得了
药家鑫辩护律师路刚	赢得官司，挽救当事人，获取更大声誉	当事人得到最严厉惩罚，最终伏法，路刚的激情杀人论广为天下人诟病。
李玫瑾等砖家精英	专业解读能得到广大观众的理解和敬服，赢得更强声誉	被千万人唾弃，有委屈成分；从专业的角度来看，有些言论不失其水准，但是其错误时机与错误传播方式铸就其专业不容理解，并被骂名无法解脱。
CCTV等代表药家鑫立场的媒体	得到好的口碑，树立媒体的权威性和公正性。	前期过于露骨地干涉法律公正，国家喉舌的威信扫地
药家鑫背后的公关团队	危机公关显身手	前期密集的危机公关适得其反，引起全民公愤，药家鑫丢命又丢脸的下场将公关团队定在耻辱柱上
广大义愤的网友	匡扶社会公平正义，是罪大恶极者伏法，与社会邪恶势力斗争	前期被药家鑫危机公关激怒，发起了全民维权运动，有一定积极意义，但后期逐步被张妙律师及一些别有用心的人变相操控，后期，药家鑫方面的声音基本沉默，最后逐步转为舆论暴力，挖隐私谩骂无辜。药家鑫毕竟还不完全是十恶不赦的恶魔，后期的遗愿以及部分作为还是有其可悲可叹可怜之处。
法院政府等公权力机关	树立国家机关的威信，为群众所拥	法院投票调查、迟迟不予以响应以及配合央视报道的种种行为导致政府公

旅游营销案例：天仙妹妹

↗：阿坝州旅游区之“天仙妹妹”传播方案

营销背景：阿坝州旅游区是一个风景秀丽的地方，是中国著名的旅游景点之一，包括九寨沟、花湖、羌寨等，但是知名度很低，导致在旅游旺季客流量都很少，提高阿坝州旅游区的知名度迫在眉睫。

阿坝州旅游区的优势：风景如画、姑娘漂亮、度假旅游胜地。
策略创意：将阿坝州旅游区局部放大进行传播，网友对单纯的广告不感兴趣，所以需要通过感兴趣的点进行结合。

广告效益：

网易、新浪、搜狐等门户网站均开辟天仙妹妹专题

- 传统媒体跟踪报道，央视的《社会记录》《新闻会客厅》做专题报道
- 网络搜索量超过百万，迄今为止最为成功的网络造星案例
- 事件从2005年8月持续至今 仍不断有新闻热点
- 直接给阿坝州的旅游经济带来30%的增长

旅游营销案例：天仙妹妹

浪迹羌寨 单车川藏自驾游之:惊见天仙mm?!(转载)

早年经历 演艺经历 主要作品 公益活动 获奖记录 更多>>

baike.baidu.com/

"天仙妹妹"甜樱桃畅销,带动阿坝州旅游业振兴(组图)_网易新闻中心

2011年6月9日 - 随着"天仙妹妹"牌甜樱桃在各地打响名声,阿坝州旅游区也被越来越多的人熟知,在提高甜樱桃销量的同时,反过来促进了阿坝州旅游经济的发展。接下来,赵华和...

news.163.com/11/0609/1... V3 - 百度快照

"天仙妹妹"甜樱桃畅销,带动阿坝州旅游业振兴(组图)_...新浪博客

2011年6月22日 - 阿坝州旅游区形象代言人天仙妹妹,及其品牌持有人赵华联系到茂县林业局,表示愿意公益赞助茂县的甜樱桃产业。...

blog.sina.com.cn/s/blo... - 百度快照 - 88%好评

【映秀花开】樱桃园里巧遇羌族"天仙妹妹"-搜狐旅游

2015年5月29日 - 旅游情报爱旅行的小卡搜狐旅游 > 国内游 > 四川 > 阿坝州 > 避暑 ...不知道还有多少人记得,几年前红遍网络的"天仙妹妹"?便是因为身着羌族服饰,...

travel.sohu.com/201505... V3 - 百度快照

贾君鹏、天仙妹妹成浮云,网络公关炒作风光不再_浪潮求生_读书_...

著名的天仙妹妹也是一次网络炒作,旨在推广阿坝州的旅游资源确实曾经有一些著名的网络事件火爆网络,天仙妹妹便是其中很有名的一例。当年天仙妹妹以其清秀的容貌、淳朴的...

data.book.hexun.com/ch... - 百度快照 - 95%好评

天仙妹妹_生平_电影网

天仙妹妹生平介绍,星路历程,教育背景,人物传记等详细资料尽在电影网M1905.com。... 处家"的成都网友(本名:杨军)独自驱车在四川阿坝州旅游时深入羌寨,路上偶遇天...

原作者：浪迹天

周五晚，本

俪相聚亦甚欢，

意。什么盛夏雨

散散心也好，

行途中刚进阿坝

、“玖玖四”伉

起西行藏区之

别都市的喧嚣，

实在没想到，此

楼主发言：2次 发图：0

藏 | 更多 | 楼主 回复

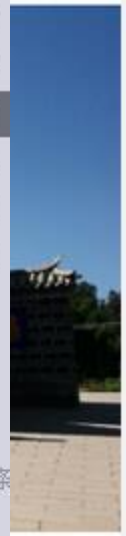
打赏楼主：

大数据分

工大學
OF TECHNOLOGY

危机公关逆转：会理PS案

危机公关逆转：会理PS案



-----社会

机---网销

大数据背景下的新媒体传播

今年实验室毕业的学生都有了好去处，百度的offer四个人，亚马逊一个，我军某部一个，二十万起薪。而且还在挑，让我们这些十万年薪的老师怎么活啊，强烈要求教育部出台规定：毕业生每年上交一月薪水给老师，连续三年，估计教育质量就上去了。 北京理工大学信息...

11月26日 17:29 来自 三星Galaxy NOTE III

阅读 34.7万 推广 转发 297 评论 6

@程序员邹欣 你是个坏人，我要是家里揭不开锅，我去你们微软要饭去 另外，更正一下，我军待遇没那么高，都是按照规定做的，但前途比钱途好啊。 //@程序员邹欣:同意，找不到工作的由老师亲自负责半年的工资！这样只需两年即可提高教育质量。

我们明年才正式毕业的硕士生都有了一个相对不错的归宿，这么多人互动。头一回微博比微信活跃。理工大学的意义在先从能养活自己和家庭开始 //@TJUReyoung:如果上大学学的悲哀。

@ICTCLAS张华平博士

今年实验室毕业的学生都有了好去处，百度的offer四个人，亚马逊一个，我军某部一个，二十万起薪。而且还在挑，让我们这些十万年薪的老师怎么活啊，强烈要求教育部出台规定：毕业生每年上交一月薪水给老师，连续三年，估计教育质量就上去了。 北京理工大学信息...

11月26日 17:29 来自 三星Galaxy NOTE III 阅读 34.7万 转发 297 评论 67 40

11月26日 20:10 来自 微博 weibo.com

阅读 8880 推广 转发 3 评论 1 3

梁总，你导师一支持，你就跟进表态，这是什么节奏，我这个提议教育部批准的概率很低，先从你们师生俩开始实践吧，你博士毕业三年内，我都会来监督你交租的，等着吧！ @马少平THU 收款后记得请我吃饭啊。 //@梁斌penny:支持！ //@马少平THU: 支持！

@ICTCLAS张华平博士

今年实验室毕业的学生都有了好去处，百度的offer四个人，亚马逊一个，我军某部一个，二十万起薪。而且还在挑，让我们这些十万年薪的老师怎么活啊，强烈要求教育部出台规定：毕业生每年上交一月薪水给老师，连续三年，估计教育质量就上去了。

11月26日 17:29 来自 三星Galaxy NOTE III 阅读 34.7万 转发 297 评论 67 40

北京理工大學
BEIJING INSTITUTE OF TECHNOLOGY

舆情攻防微革命

↗以“舆情源”为中心的全周期攻防

- **舆情发生期-舆情当事人**
  - 国家安全危害分子：台独、藏独、疆独、民运、邪教、暴恐；
  - 社会舆情高发群体：拆迁上访、转业军人安置、房价、就业、就医、反腐、传销经济诈骗、反社会伦理；
  - 高敏感公立群体：政府机关、官办协会、高校、事业单位
- **舆情传播期-网络受众**
  - 有影响权威大V：左派、右派、新左派、民主派、律师、公知、高级黑（各有分工的权威领域，各有特色；针对性处理）；
  - 愤青；
  - 理性质疑者
  - 沉默的大多数，沉默者的狂欢就是舆情的顶峰

新媒体舆论场派别

- 社交网络舆论场，沉默的大多数，民意主要是极左极右势力的角力场。伴随着各种转世党的角逐
  - 右派： @陈有西；@陈志武；@大鹏看天下；@高会民；@贺卫方；@胡紫微；@克里斯托夫-金；@李悔之2012；@李剑芒的小号；@李开复；@慕容雪村；@诗人潘婷；@孙君红 ；@吴稼祥；@吴祚来；@夏业良七世；@信力建；@徐昕 北理工法学教授；@薛蛮子；@袁莉wsj；@袁腾飞；@袁伟时；@袁裕来律师；@章立凡；@赵楚；@赵晓；@中青报曹林；@左小祖咒；@作业本；@茅于轼
  - 左派：@孔庆东 @司马南 中共中央政策研究室综合局局长张勤德、中央民族大学教授张宏良、中国人民大学教授贾根良、中国政法大学教授杨帆、北京航空航天大学教授韩德强、《光明日报》原副主编陈谈强、中国现代国际关系研究院经济安全研究中心主任江涌、原国史学会副秘书长苏铁山和剧作家黄纪苏
  - 新左派：杨帆@高梁@何新@旷新年@张广天@黄纪苏@胡鞍钢@韩毓海@王绍光@汪晖@黄平@崔之元@甘阳@巩献田

转世党去哪了？

userId	昵称	href	粉丝	关注	微博数
2885194051	茶马古道K	http://weibo.com/u/2885194051	1539	1973	4317
2494346694	roadkiller007	http://weibo.com/u/2494346694	881	1780	1218
3210204934	草l民l嘟l囔	http://weibo.com/u/3210204934	8356	2000	6223
3705439931	忠殃政府	http://weibo.com/u/3705439931	979	1343	1525
3843158184	别开枪我是逗比	http://weibo.com/u/3843158184	1469	683	573
2494346694	roadkiller007	http://weibo.com/u/2494346694			
3636801343	草原狼ZG2	http://weibo.com/u/3636801343	1211	1999	4000
3786413383	恶心的花	http://weibo.com/u/3786413383	2821	1869	5397
3579344034	二黑媳妇III	http://weibo.com/u/3579344034	4168	888	2575
5405551302	永苗在成都	http://weibo.com/u/5405551302	1679	1333	487
3907742931	挑燈買醉	http://weibo.com/u/3907742931	1984	461	73
5339858836	草长老x	http://weibo.com/u/5339858836	897	913	1288
5210911387	率率妮子	http://weibo.com/u/5210911387	3358	963	4742
1766678473	心随风动111	http://weibo.com/u/1766678473	4326	827	13784
1712838332	530涅槃重生14	http://weibo.com/u/1712838332	661	294	355
5547246671	龙逸天541	http://weibo.com/u/5547246671	566	379	2
5202504558	乌托国人民	http://weibo.com/u/5202504558	1123	1101	11193
5556809166	小军的自油18	http://weibo.com/u/5556809166	236	302	115
2700138820	李不白的微博	http://weibo.com/u/2700138820	72818	791	12864
	新闻已死	http://weibo.com/thenewshasdied			

转世党去哪了？

userId	keywords											
1712838332	重生	涅槃	变成	老罗	微评	李不白	爱卿	免礼	青山	冰哥	平身	报人
1766678473	极权	古越	心随	民主	风动	自由	权利	阿宝	权力	专制	思维	王兰墨
2494346694	三十而立	safiya	律师	徐昕	roadkille	卧槽	元芳	中国	咳咳	迟夙生	袁裕来	反腐
2700138820	翁涛	小店	天佑	浩正	刘臻	律师	呵呵	贺江	国家	亚军	中国	雾满
2885194051	轉發	律师	礼江	美鱼	中国	彭园	枫叶	秦时	aaa	流氓	国家	嘉佑
3210204934	律师	中国	自油	浪子	小军	维权	杨鸿	反腐	腾讯	政府	关注	浩正
3579344034	刘植荣	警察	公务员	中国	乌克兰	公示	反腐	美国	越南	民富国强	财产	罢工
3636801343	ZG	草原	沙鸥	文革	律师	徐子升	中国	美裙	东平	岁月	醉侠	猫咪
3705439931	律师	赵玉敏	共产党	代新红	徐昕	天津	大姐	神评	正义	玉麒麟	中国	陈胜德
3786413383	家人	寒冰	幼稚	律师	迟夙生	绝世佳人	醉侠	孙海英	老高	鼠标	萌军	善良
3843158184	律师	哈哈哈	醉侠	老高	九世	文三娃	范木根	红包	美猴	老王	石扉	智勇
3907742931	谈史	司马	故乡	勤政	子女	哈哈哈	孙海英	富時	彗心	張靈甫	梁啟	中國人
5202504558	律师	砀山	微博颖王	李不白	自油	卡扎菲	腾讯	小军	泥人	照新宇	人要	徐昕
5210911387	率率	妮子	真妮花	AV	美帝	大总统	乖乖	花花	老頭	五毛	微评	流浪
5339858836	极权	推享	莫大	古越	先生	salavivo	公知	政治	民主	社会	体制	自由
5405551302	民国	改革	体制	共党	大陆	知识分子	美国	国体	当归	政治	台湾	台独
5547246671	邪恶	天朝	秘书	天渊	血债累累	暴毙	力加	国家	哪出	睡马大	反烟	党旗

转世党去哪了？

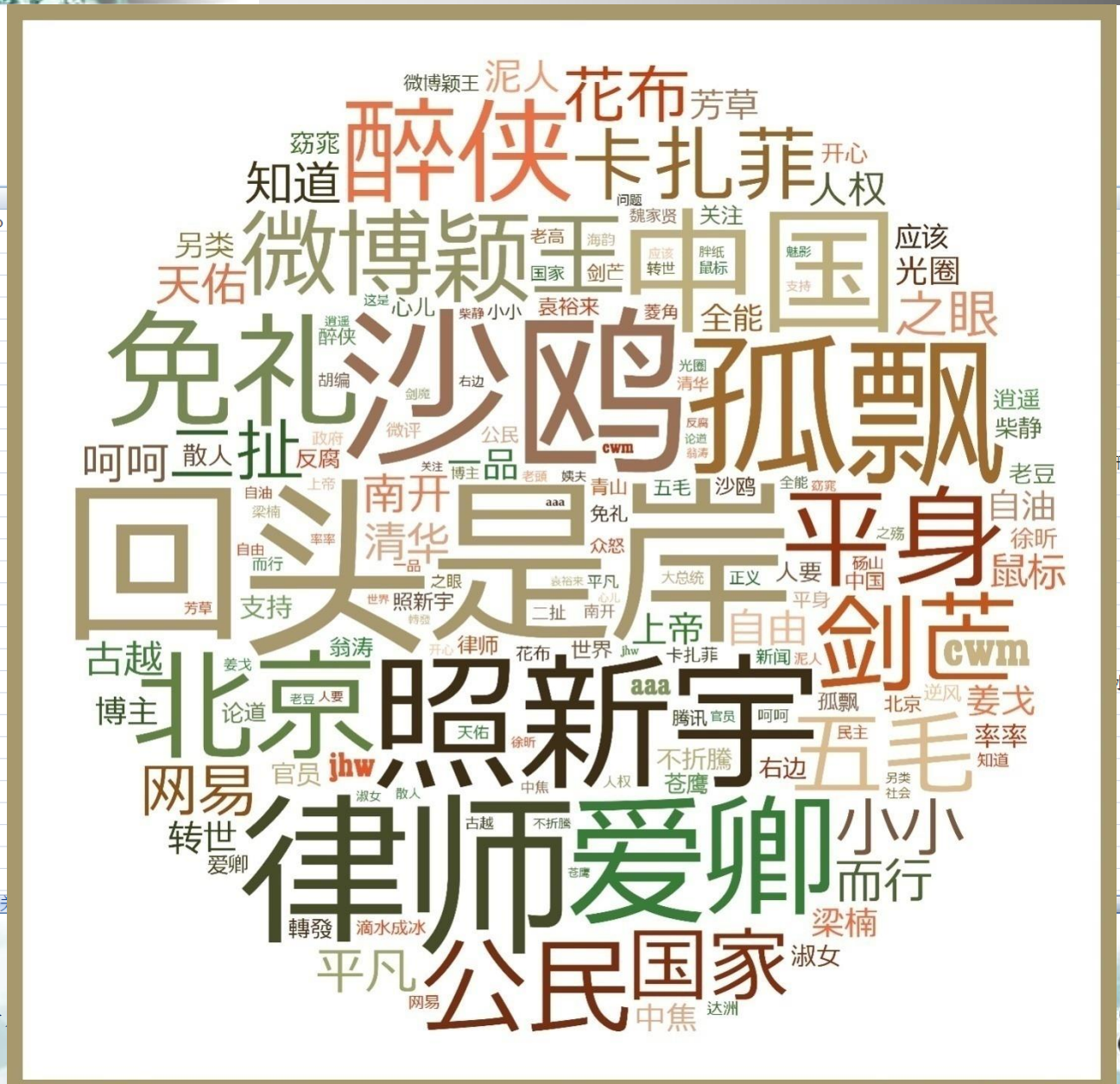
转世党去哪了？

- 依群体1关注者信息及群体特点，扩展群体名单
- 新增4000人
- 新增群体大部分与律师、民主、攻击党和政府等话题相关

转世党去哪了？

- 抽取扩展名单45人微博内容，进行关键词分析，定义匹配系数（matching_ratio），即扩展人员微博内容与给定名单话题相关比例
- 匹配系数在30%以上有18人，约占40%

转世党去哪了？



	A	B
	userid	matching_ratio
1	5498180920	0.7
2	5202496301	0.55
3	2692425403	0.45
4	5262631069	0.4
5	5458304288	0.4
6	5545484709	0.4
7	1203733845	0.35
8	5314732543	0.35
9	5318216818	0.35
10	5420262586	0.35
11	5446215890	0.35
12	1458137814	0.3
13	1662245660	0.3
14	2445716161	0.3
15	3208475461	0.3
16	5319608938	0.3
17	5321242214	0.3
18	5425985422	0.3
19	2799024532	0.25
20	2970946642	0.25
21	5138757880	0.25
22	5407787865	0.25
23	2655170343	0.2
24	3977808825	0.2
25	5230807804	0.2
26	5262287792	0.2
27	5459116340	0.2
28	1671685587	0.15
29	1732403974	0.15
30	5149052923	0.15

基本信息 关键词 全部微博

R	S	T	U
美鱼	jhw	戎戎	财产
逆风	腐败	国家	北京
老百姓	暴力	政府	马航
他妈的	回头是岸	滴水成冰	北京
微信	维稳	选票	共产
律师	众怒	自由	爱卿
官员	老豆	潜龙	张晨
事儿	命题	尼玛	独裁
知识	素质	革命	柴片
中焦	免礼	渔夫	天佑
立案	正义	芳草	六安
张佳	公知	喜水爱山	团练
五毛	两会	网铂程	时评
AM	包子	支那	逆风
不加	童鞋	老頭	秀才
aaa	妩媚	论道	真妮
鼠标	三世	中国	盼春
宠爱	美好生活	内河	林下
榨取	社会	架构	意识
邓相超	周老川	照新宇	文章
腾讯	卡扎菲	二扯	老豆
北京	金枪鱼	逆风	滴水
律师	牲口	菲尔	光圈
老齊	徐昕	花布	日记
畜牲	清华	菱角	落拓
海韵	卡扎菲	姜戈	魏家
孤飘	二扯	花布	律师
政府	千八百	朋友	时事
六老	Fujimi	砀山	苍鹰
徐昕	故现	贺卫方	袁立

大数据分

工大學
OF TECHNOLOGY

纲要

大数据
新媒体

I 新媒体传播创新

II 社会化新媒体舆情特点

III 社交媒体分析关键技术

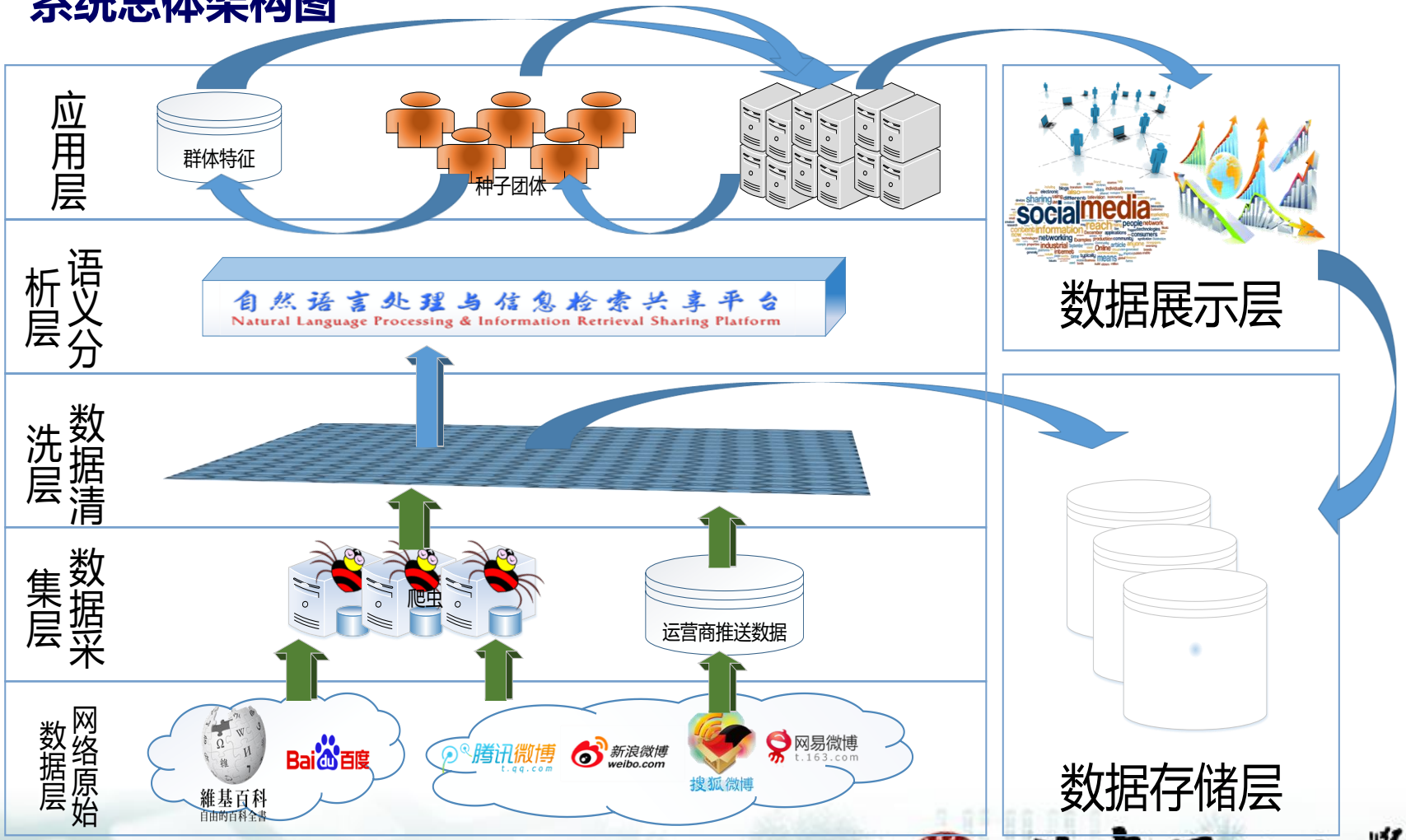
IV 大数据新媒体应用实战案例

研究内容

- 研究并实现特定群体与敏感用户社交网络账号的发现；
- 研究并实现对特定群体及敏感用户的社交属性、活动属性、位置属性等的全特征计算；
- 实现针对特定小众化群体及人物的快速搜索、关联分析和属性标注；
- 对特定小众化群体及人物进行快速搜索、关联分析和属性标注；
- 对特定事件或特定话题参与人员进行搜索和关联，分析参与人员在事件中或话题传播过程中的作用，为识别事件的幕后推手提供决策支持；
- 针对已构建的特定群体及敏感用户，能够查看其社交属性、活动属性、位置属性等，初步实现对突发事件的预警预报及发展态势研判；

技术路线——系统架构

系统总体架构图



北京理工大學
BEIJING INSTITUTE OF TECHNOLOGY

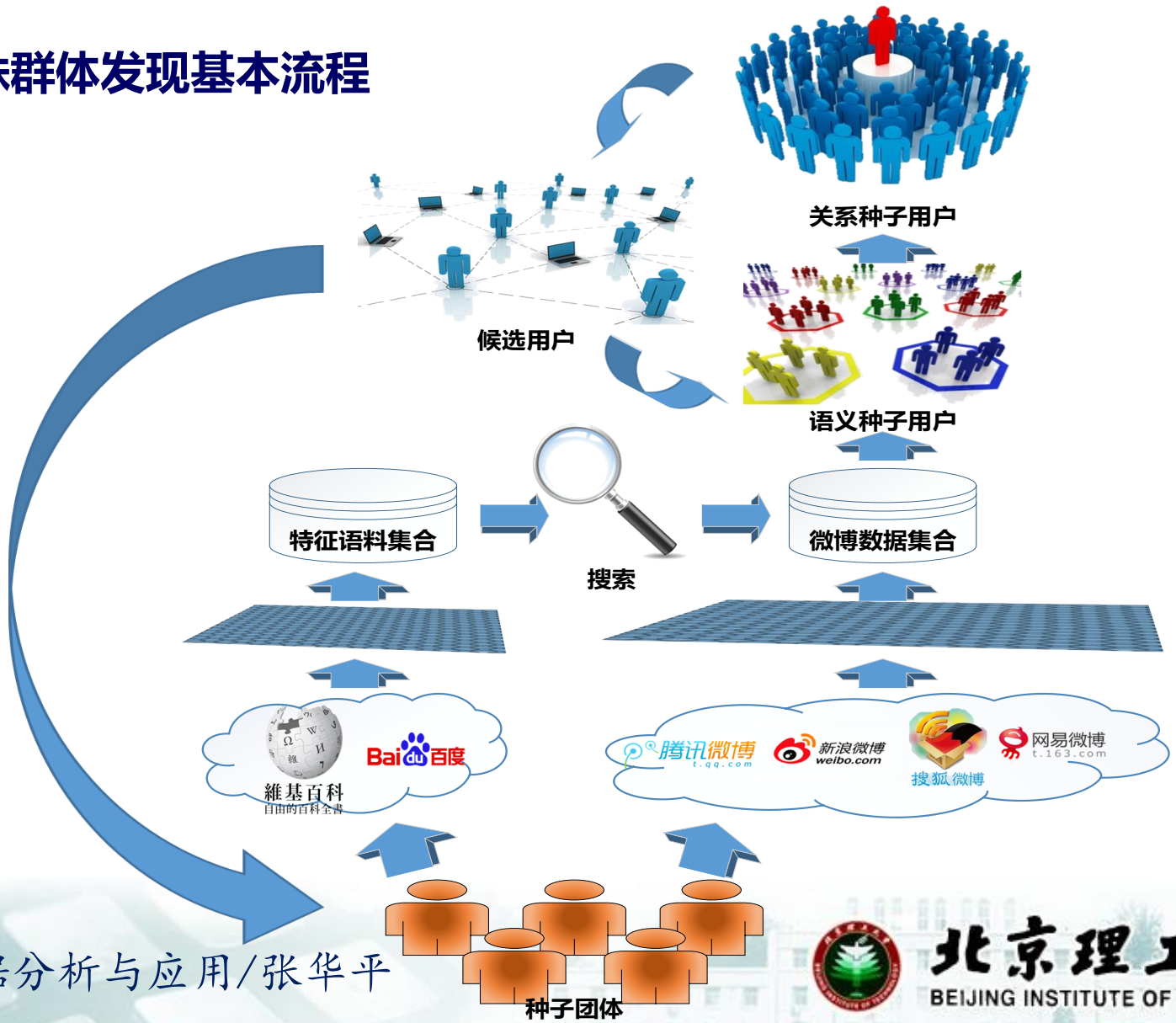
技术路线——系统架构

- **数据搜集层**
  - 中心现有数据源；
  - 配以互联网数据采集作为补充。
- **数据清理层**
  - 利用中心大数据平台的语料清洗工具对数据进行清洗；
- **语义分析层**
  - 利用中心大数据平台的自然语言分析工具，可以对数据进行社交关系和语义方面的各类运算；
- **应用层（研究目标）**
  - 利用语义分析层计算的结果实现研究目标中提到的应用；
- **数据表示层**
  - 用可视化技术对应用层输出数据进行合理表示；

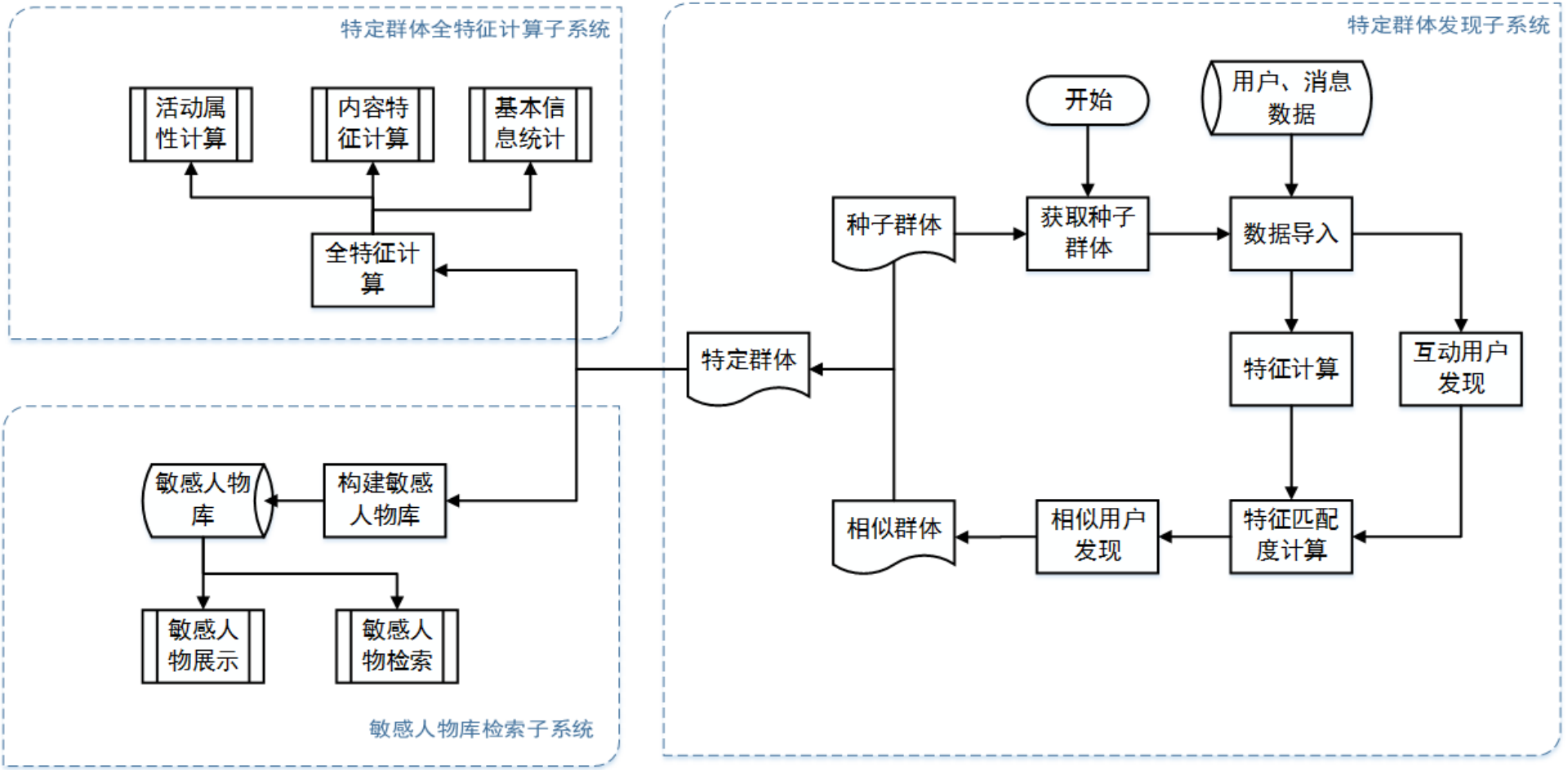
；

技术路线——特殊群体发现算法

特殊群体发现基本流程



技术方案


特定群体全特征计算子系统
活动属性计算
内容特征计算
基本信息统计
全特征计算
特定群体发现子系统
开始
用户、消息数据
种子群体
获取种子群体
数据导入
特定群体
特征计算
互动用户发现
相似群体
相似用户发现
特征匹配度计算
敏感人物库
构建敏感人物库
敏感人物展示
敏感人物检索
敏感人物库检索子系统

纲要

大数据
新媒体

I 新媒体传播创新

II 社会化新媒体舆情特点

III 社交媒体分析关键技术

IV 大数据新媒体应用实战案例

社交网络群体分析实践

- 微博大数据挖掘
- 某大V分析
- 张灵甫事件分析
- 失独老人群体跟踪

微博大数据挖掘的价值

- 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。

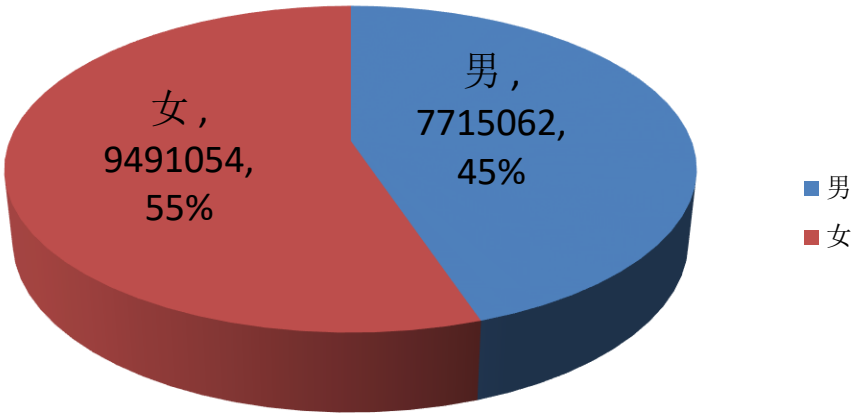
微博用户数据样本

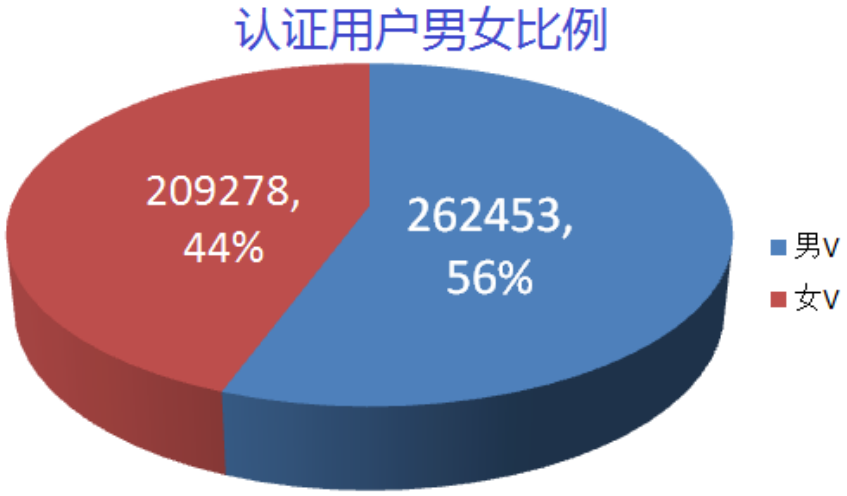
id	url	name	s..	brithday	address	fansNum	summary	wbNum	gzNum	blog	realName
10315	http://weibo.com	老军鹏	男	<NULL>	北京,海淀区	209	他还没填写个人介:	446	157	<NULL>	<NULL>
10318	http://weibo.com	王海荣	男	<NULL>	<NULL>	600	<NULL>	<NULL>	<NULL>	<NULL>	<NULL>
10362	http://weibo.com	kanglicoco	女	<NULL>	<NULL>	454	懂点人事儿的小清:	837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

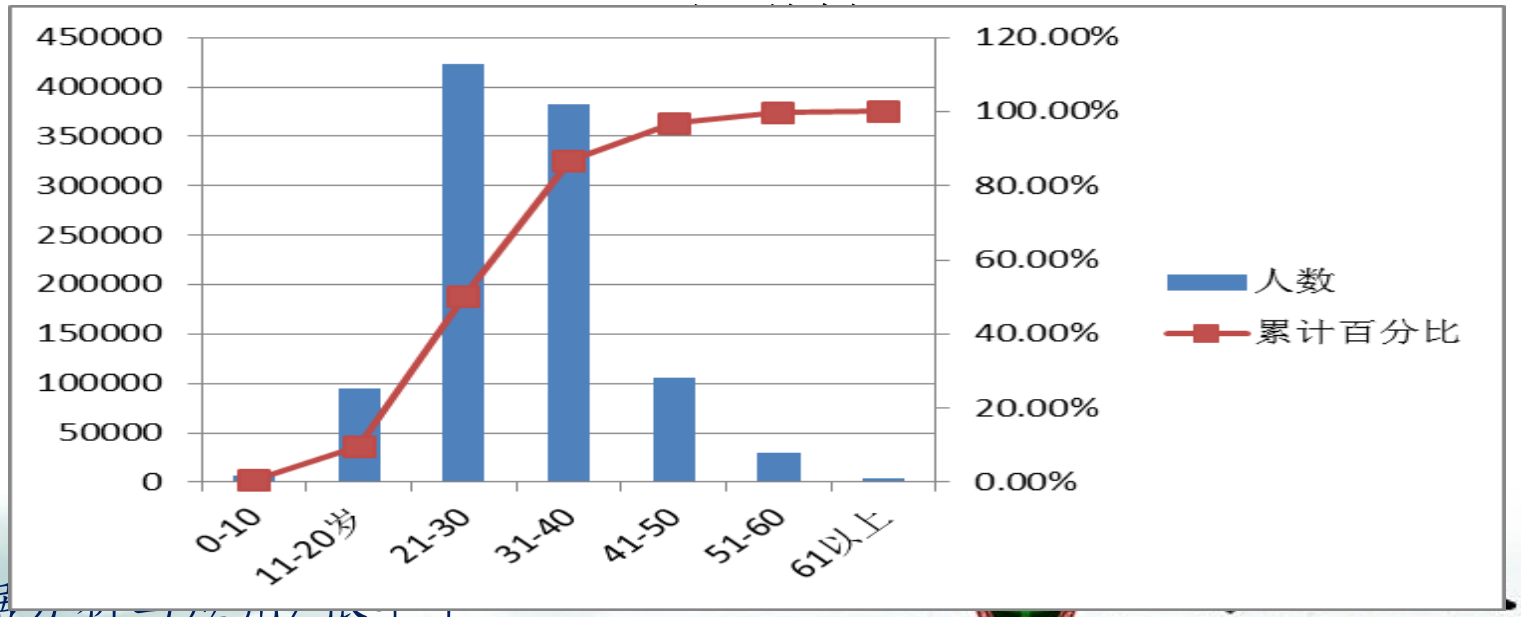
男女比例图表





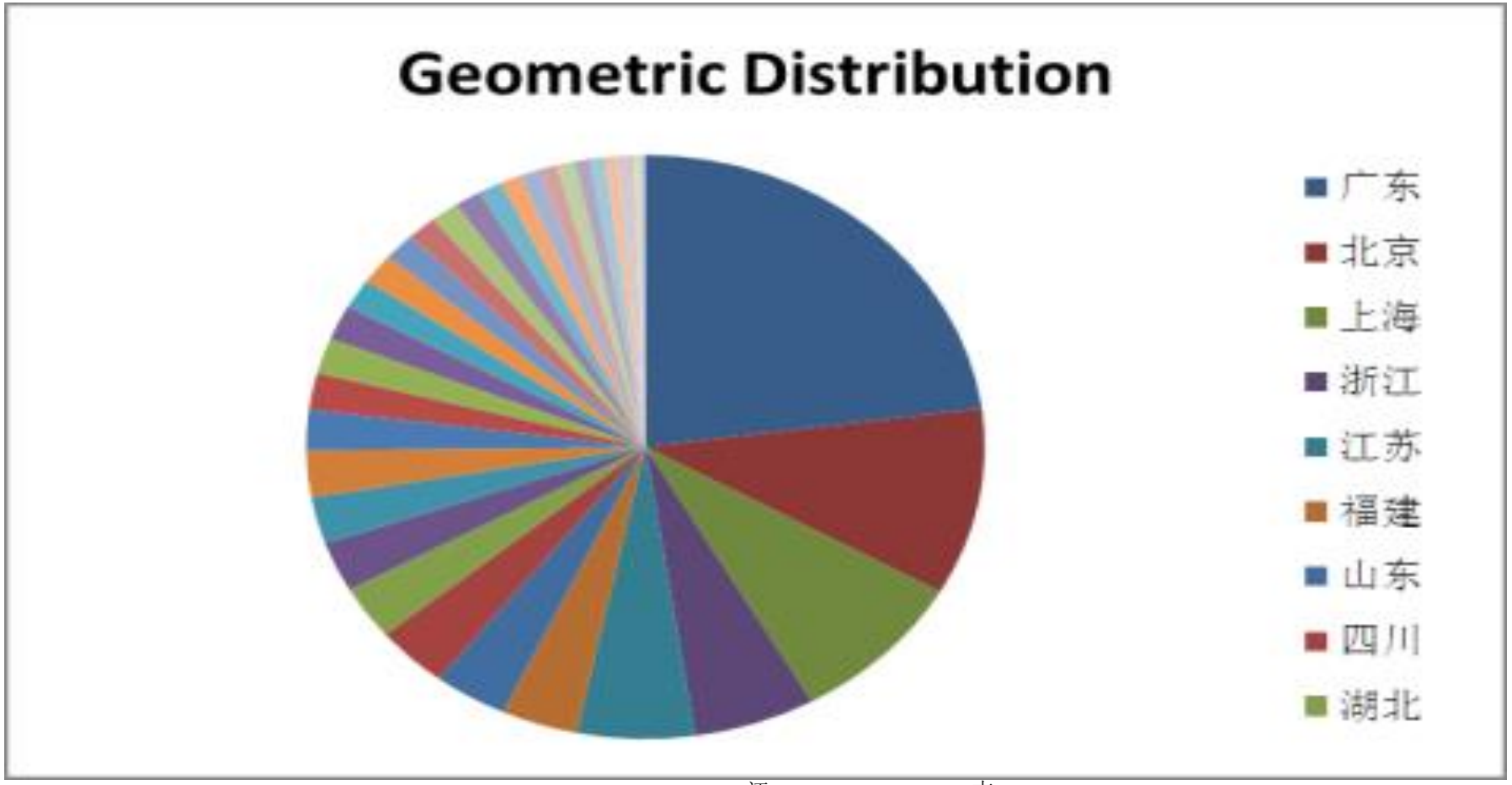
不同类型用户的分布

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

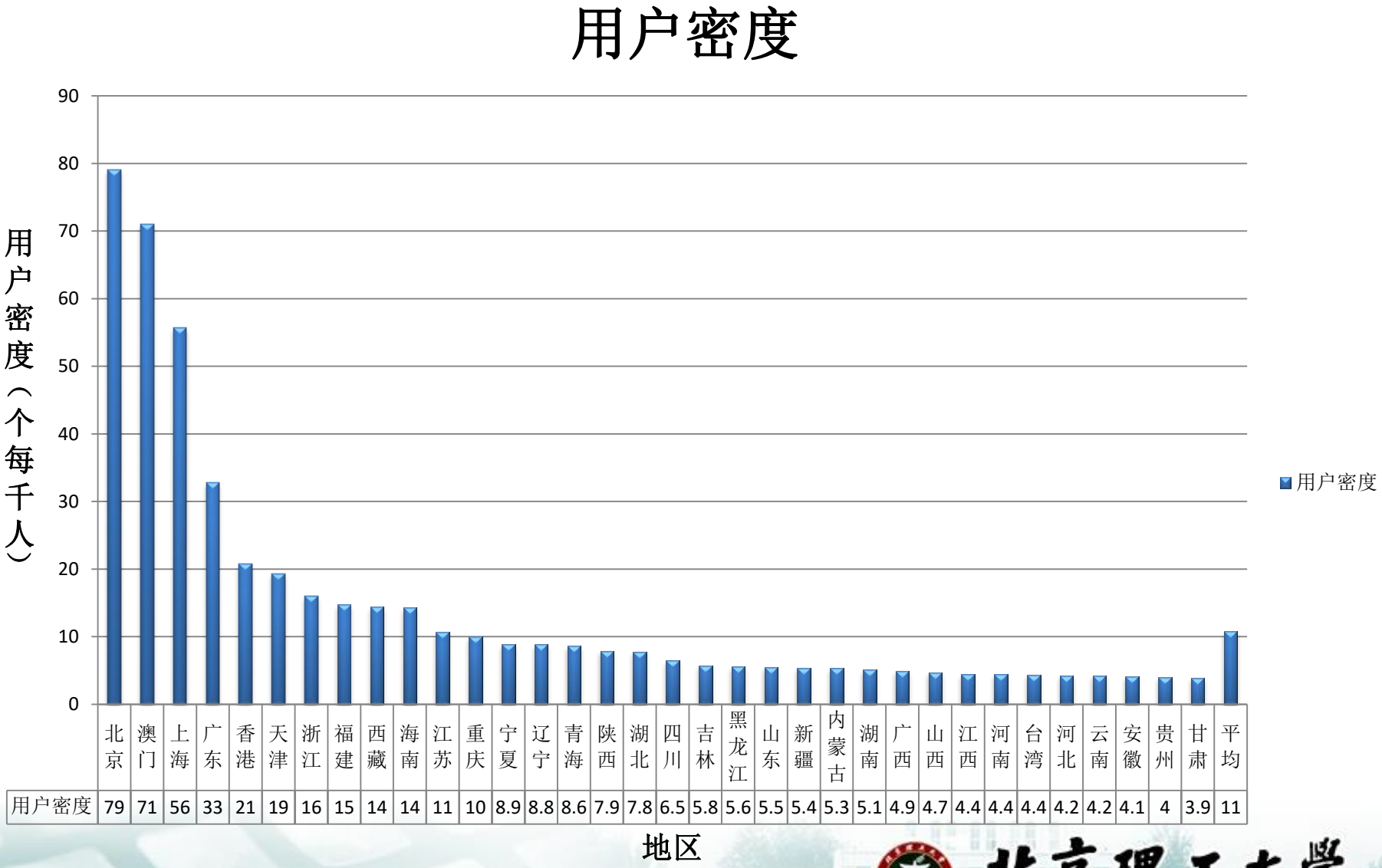
总用户数



江 古

地区

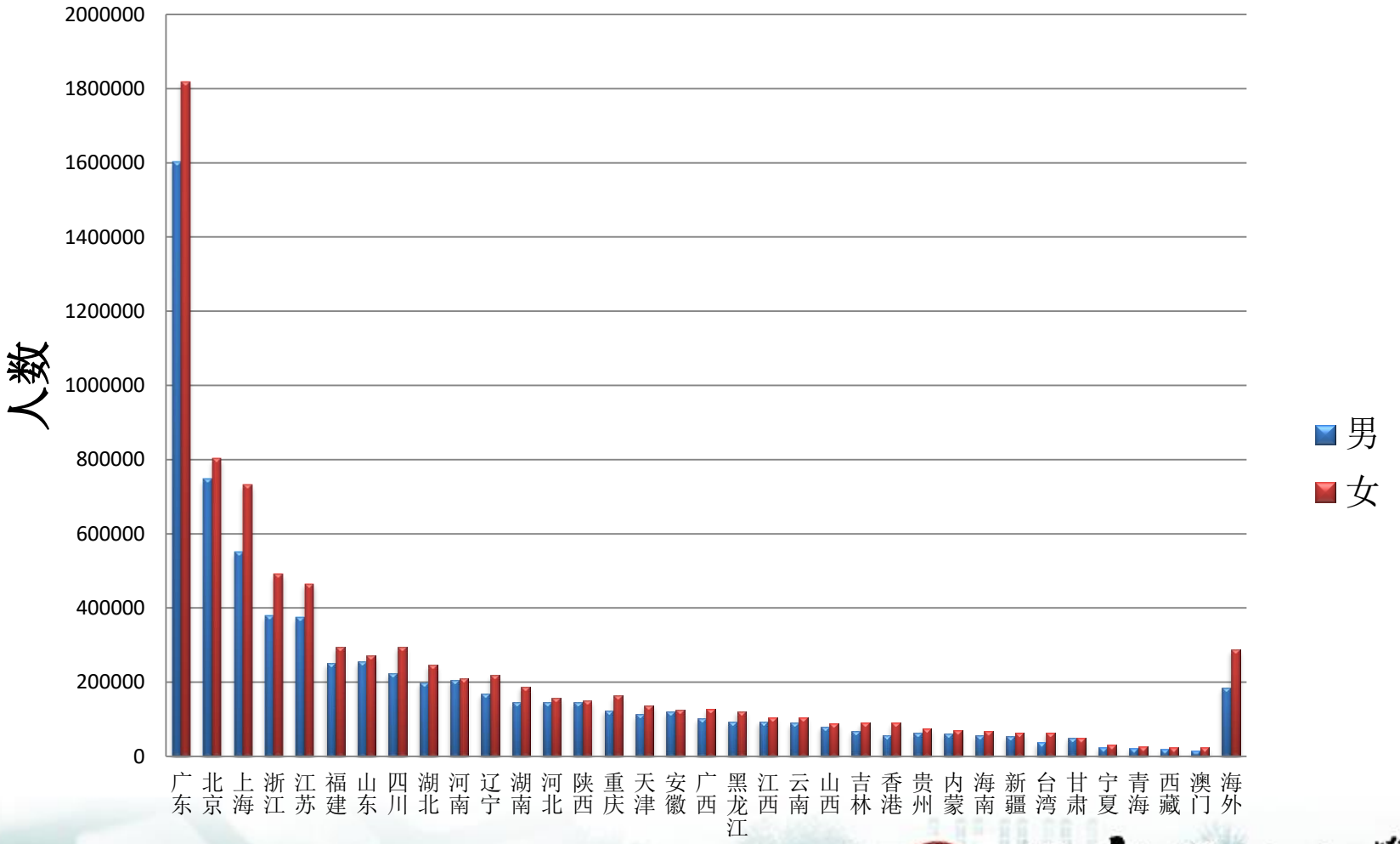
不同地区的微博用户密度



	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



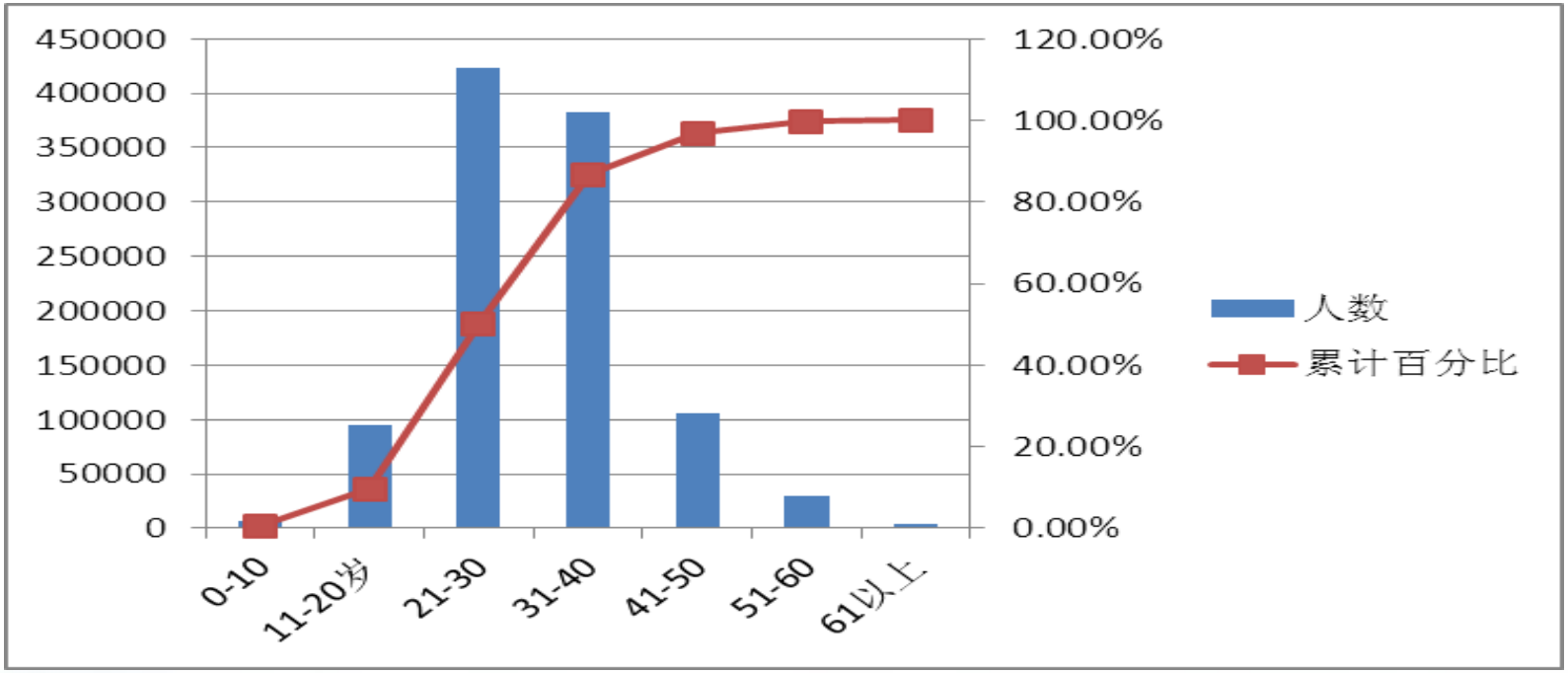
省市区划内微博用户数与GDP正相关

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

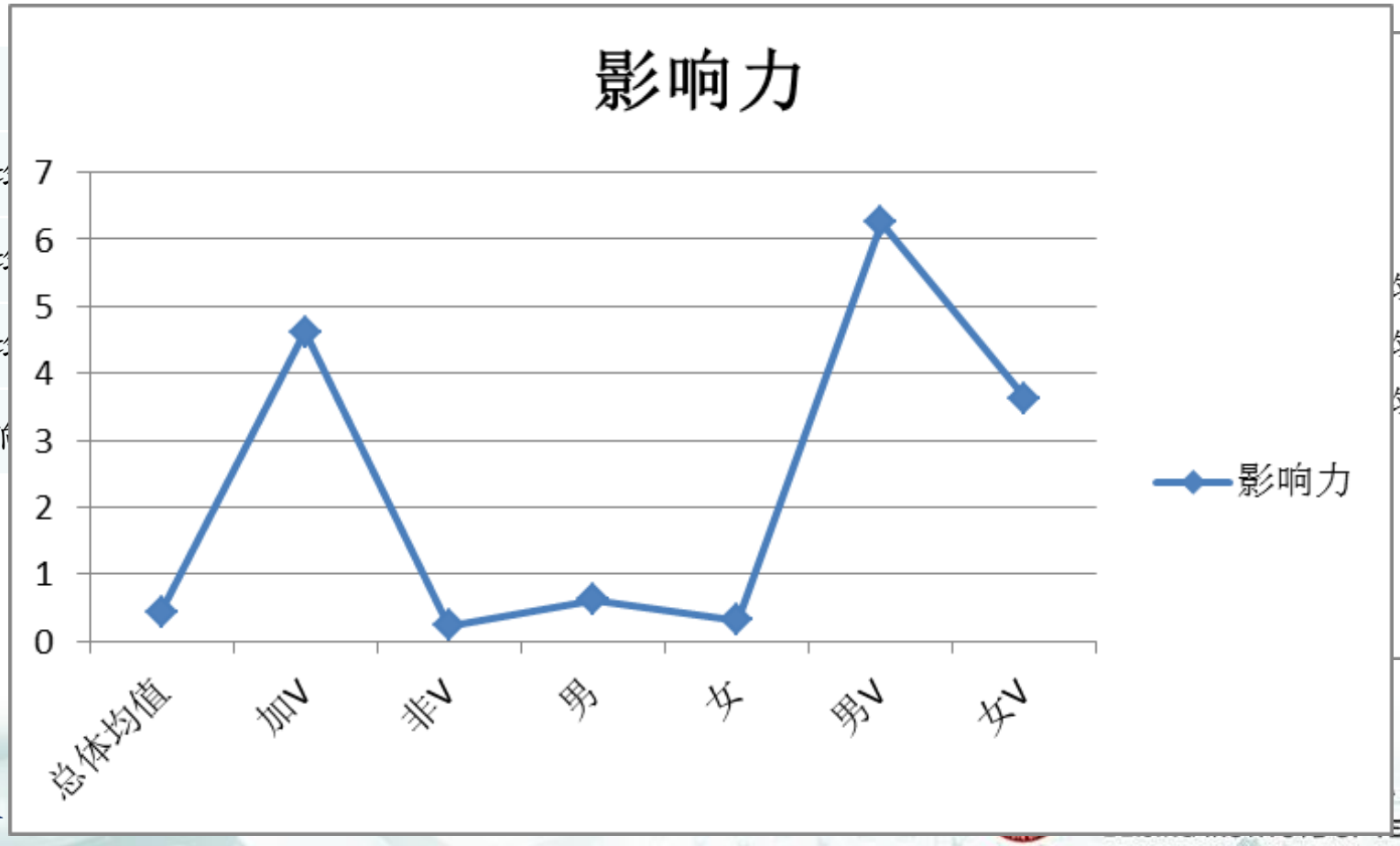
- 662,565登记了教育信息,占总人数的3.8%；其中551286大学毕业或在读,83.20%。



不同类型用户的影响力分析

影响力计算算法:

- Influence=(#fans-#following)/#tweets

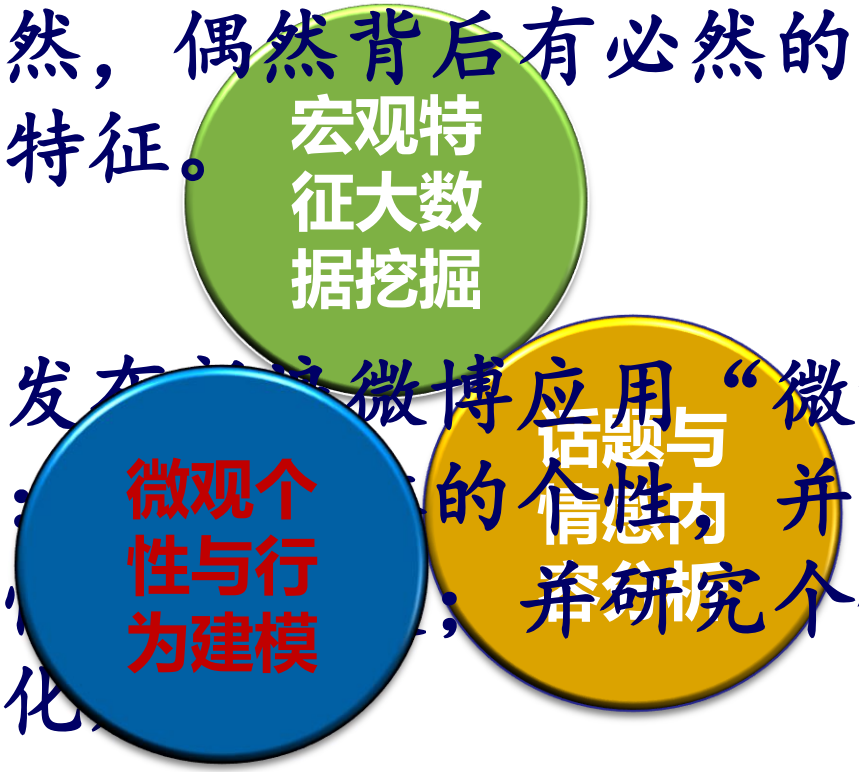


自我介绍文本挖掘

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/

博主个性化建模：沈阳教授

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

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！ |
“nlpir”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://esyrt.sinaapp.com试一试吧！

分享到微博

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	1	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	0	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

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

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%

- 加权求和?
- AHP ?
- 向量空间的欧氏距离?
- ……

大数据分析与应用/张华平

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

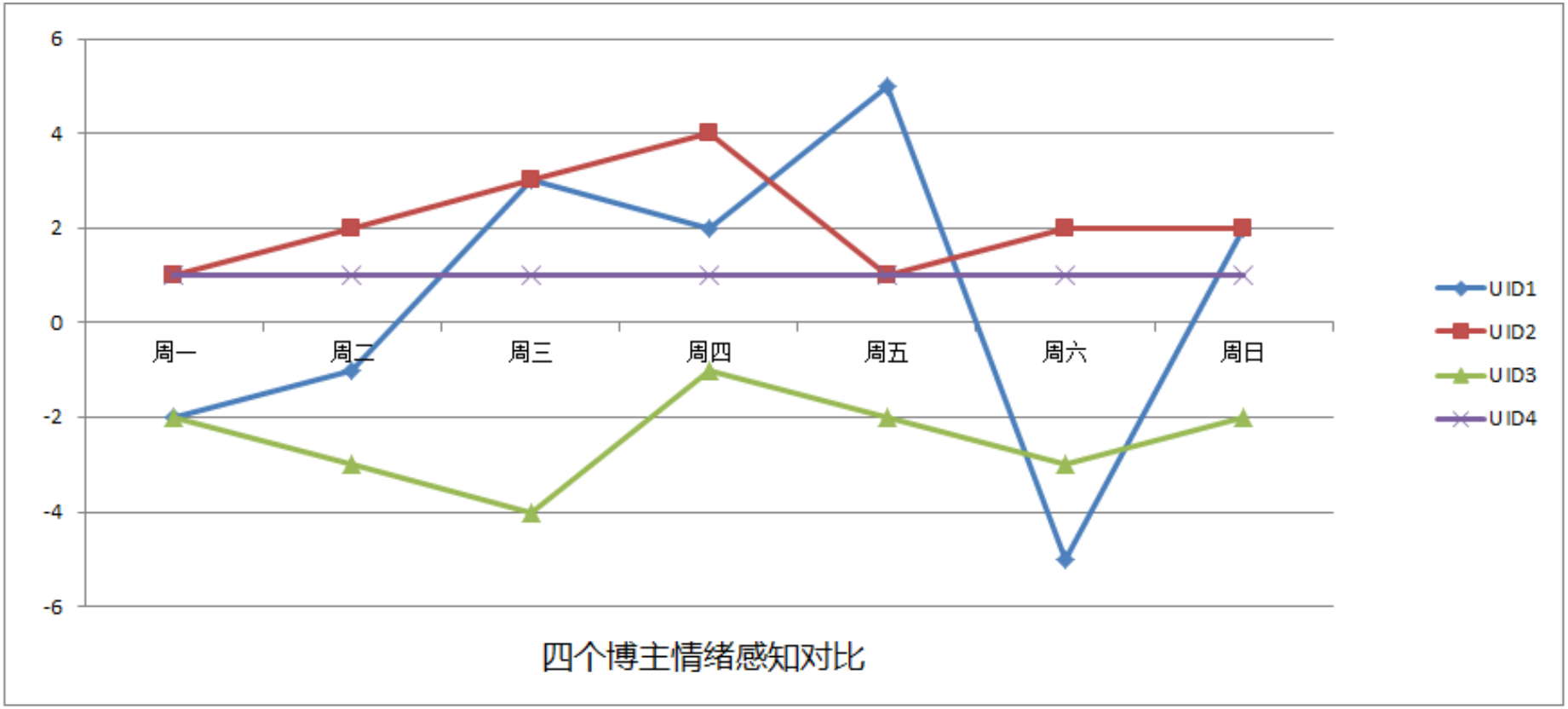
话题与情感内容分析

- 微博话题刻画
- 微博情感分析
- 感知

微博话题刻画


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

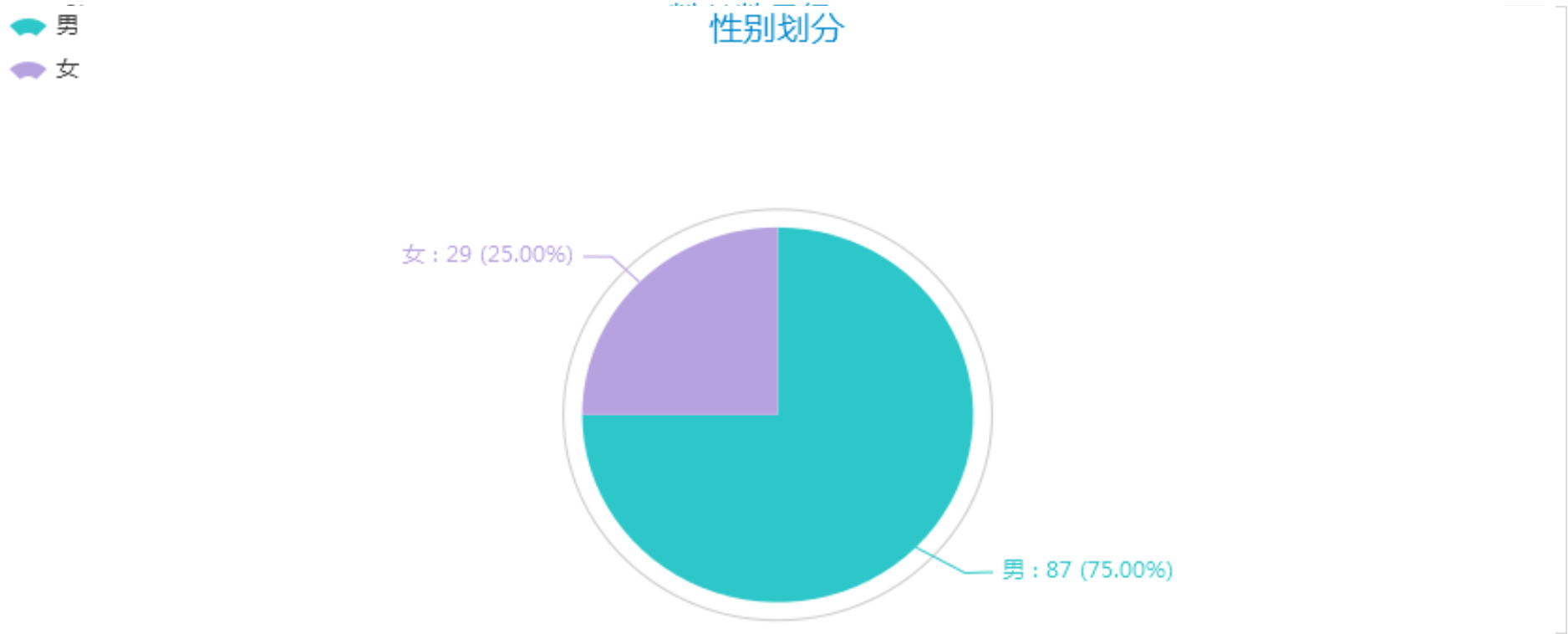
微博博主情绪感知



大数据：一言一行背后的价值观

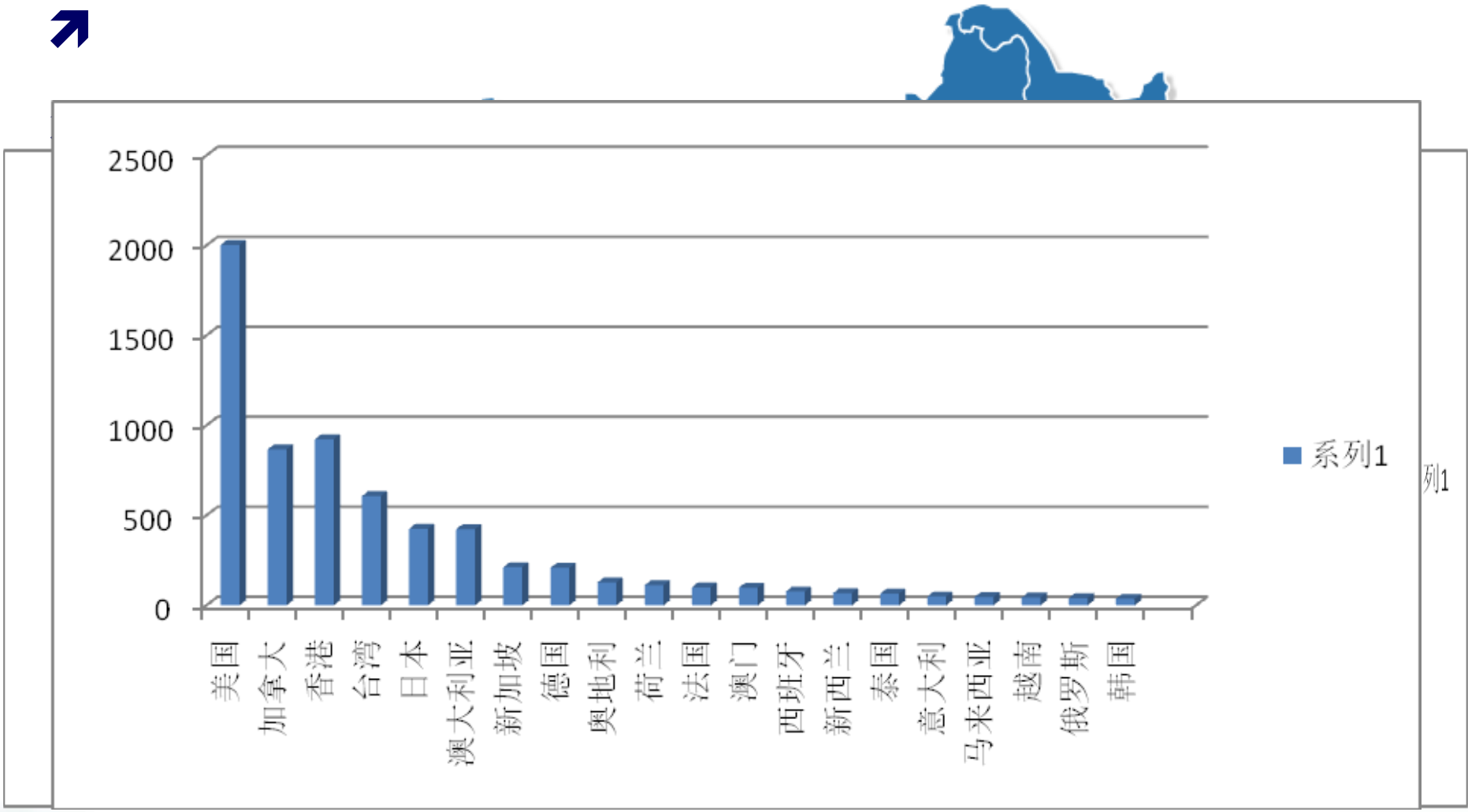
id	昵称	身份	pid	类	仁爱向量	传统向量	刺激向量	安全向量	成就向量	普世向量	权力向量	自我定向向量	遵从向量
116	孙健追求arete	国际关系学院	####	7	10	5	6	3	4	9	2	1	8
18	账号异常		####	2	10	4	5	6	3	8	7	1	9
39	IT疯云		####	2	10	1	4	7	5	8	9	6	3
53	全球热门新闻搜罗	新闻机构	####	1	10	2	4	8	6	7	9	3	5
40	海天5	基督教伯特利教	####	1	10	2	5	4	9	6	8	3	7
83	城管不好干		####	1	10	2	3	7	8	6	4	5	9
74	陈晓发	广东工业大学	####	8	10	1	3	4	9	6	5	7	2
2	喻国明	中国人民大学新	####	1	10	3	2	8	6	5	9	4	7
65	ICTCLAS张华平博士	张华平博士的微	####	1	10	3	4	8	6	5	7	9	2
5	小c是纯洁的猴子	西南石油大学二	####	9	10	2	4	6	8	5	1	3	7
68	林芙比	香港大学（200	####	6	10	2	9	3	8	5	1	4	7
103	小珍_QQ596859972	中医养生投资顾	####	2	10	1	7	8	9	5	4	3	6
95	我不叫成韦华	上海师范大学	####	5	10	3	6	1	3	4	9	8	7
19	李承鹏	记者、评论员	####	8	10	2	6	9	3	4	7	1	5
121	雙低青年	司法腐败严重则	####	5	10	1	8	6	3	4	7	2	9
78	Queen奇琦琪		####	6	10	4	9	1	5	3	2	8	8
30	东营日报社徐艺菡	东营日报社 东	####	5	10	2	6	9	8	3	1	4	7
92	Valder_	广州市商贸职业	####	4	10	1	8	6	7	3	9	5	2
62	青菜小玩子		####	6	10	3	4	9	8	2	7	1	5
70	mama咪吖	华南农业大学	####	8	10	7	9	2	5	1	3	6	4
26	何兵	中国政法大学教	####	1	9	4	2	8	6	10	5	3	7

人物分析-某大V



“张灵甫”事件的新媒体传播分析

“张灵甫”事件的新媒体传播分析



北京理工大學
BEIJING INSTITUTE OF TECHNOLOGY

所有参与者的观点分析

北京理工大學
BEIJING INSTITUTE OF TECHNOLOGY

草根的观点分析

大v的观点分析

大数据分

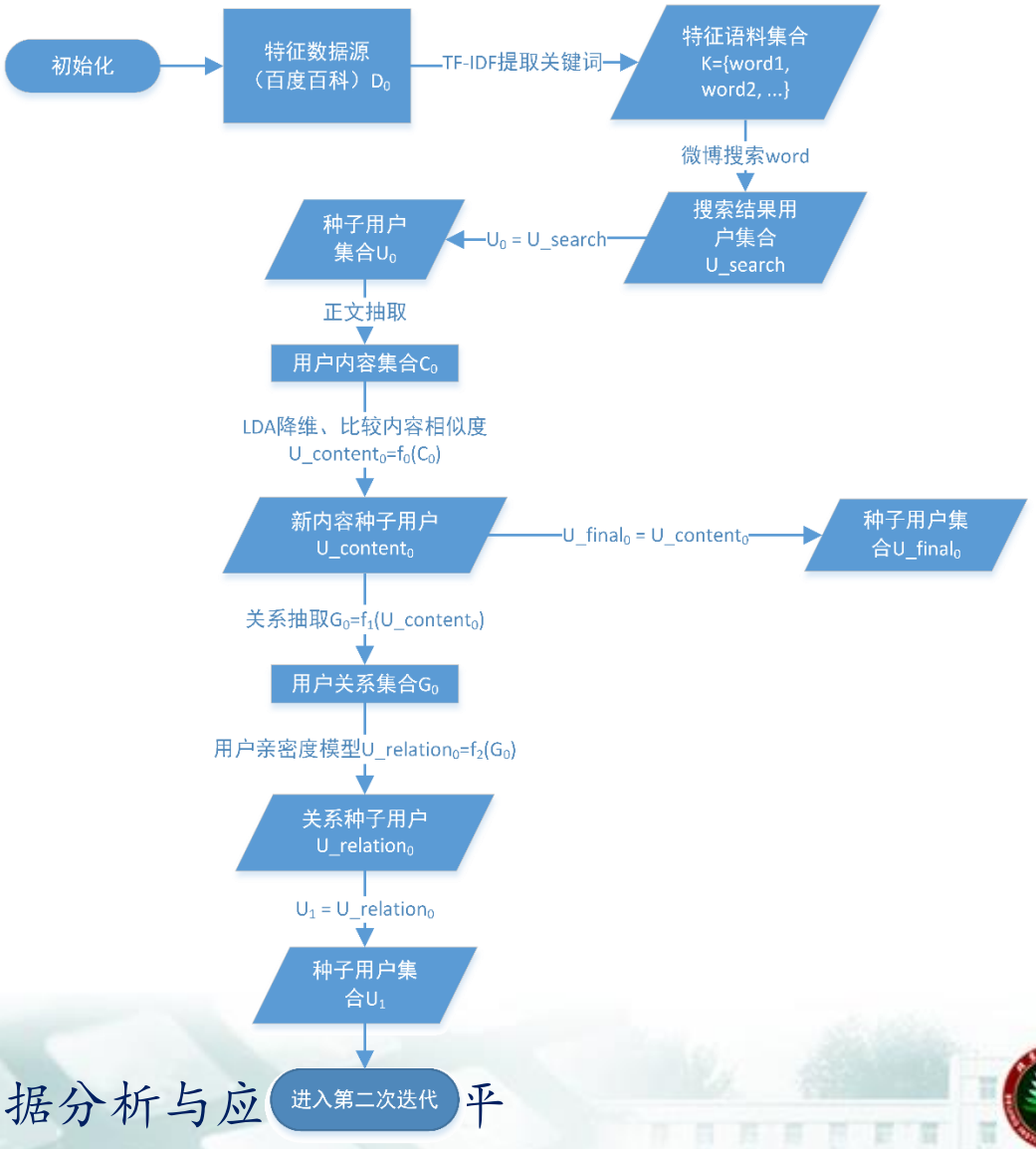
媒体的观点分析

失独老人发现项目

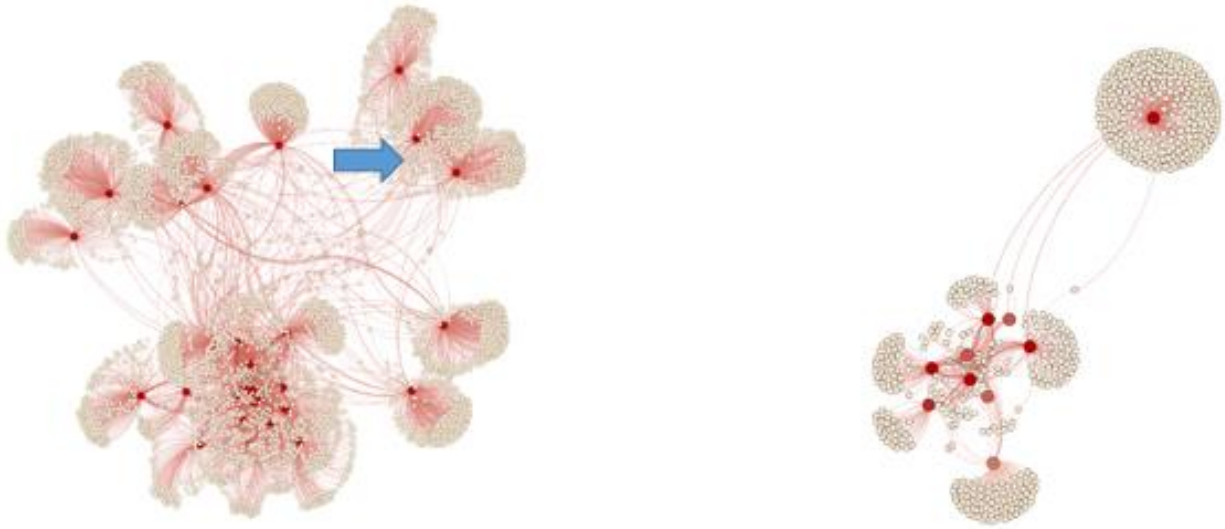
本文选择对“失独”这一主题进行实验分析。因为家中唯一的子女不幸离世，这样的家庭被称为“失独家庭”。家中的老人即被称为“失独老人”。

通过在微博平台上寻找与“失独”相关的群体，合理地检验模型的有效性，并且结合微博文本分析方法和关系分析方法对这一特定群体进行案例分析，从数据的角度对案例进行分析解释。

失独老人发现算法流程



失独老人算法迭代演变过程



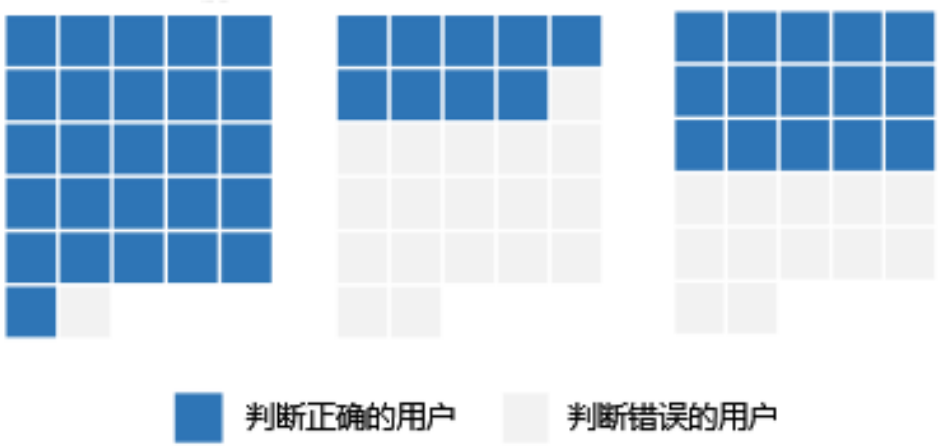
语义种子用户的关系网络的迭代演变过程

用户关系网络描述表

	第一次迭代	第二次迭代	第三次迭代	第四次迭代
节点数量	227580	12410	5490	637
边数量	296111	18536	7655	885

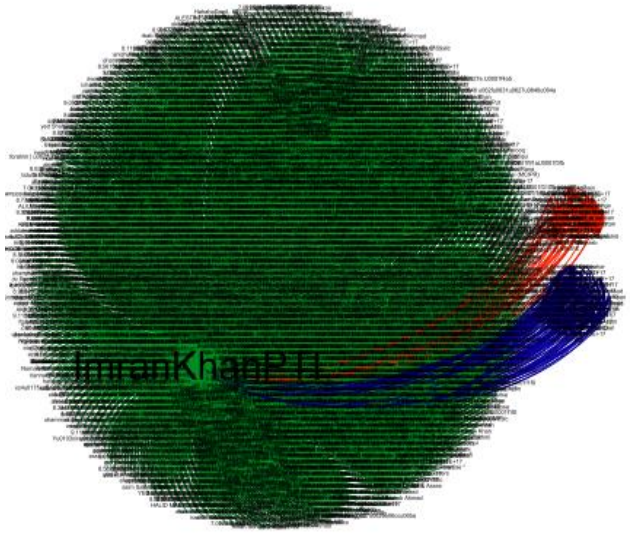
失独老人发现算法评测

	特定群体发现模型（I）	搜索发现算法（II）	改进的搜索发现算法（III）
正确的用户数量	26	9	15
正确用户占比	0.96	0.33	0.56



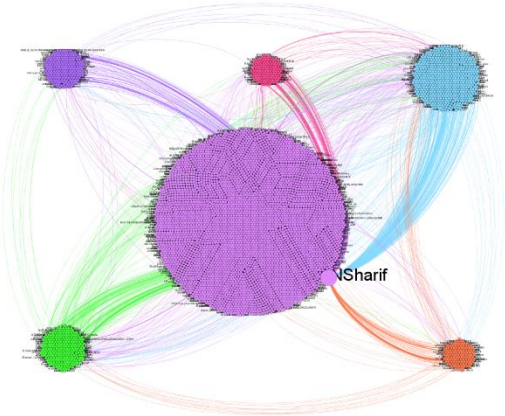
REPLY NETWORK

Imran Khan



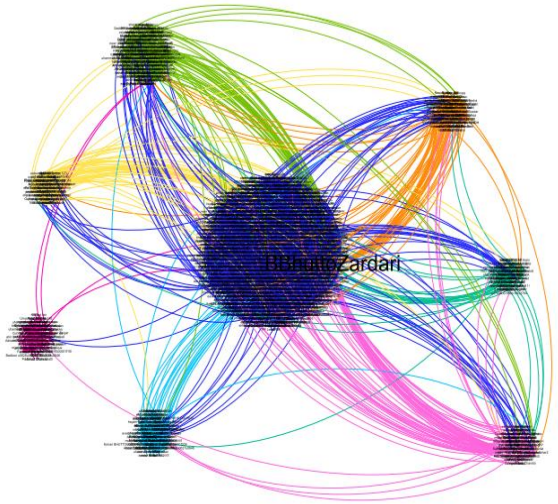
99.92% of the total network
96.47% = IK,1.97% = MN, 1.48% = BB

Maryam Nawaz



78.33% of the total
60.98% = MN, 7.71% = IK, 3.43% = other party members,
2.79% = BB, 1.71% = media cell

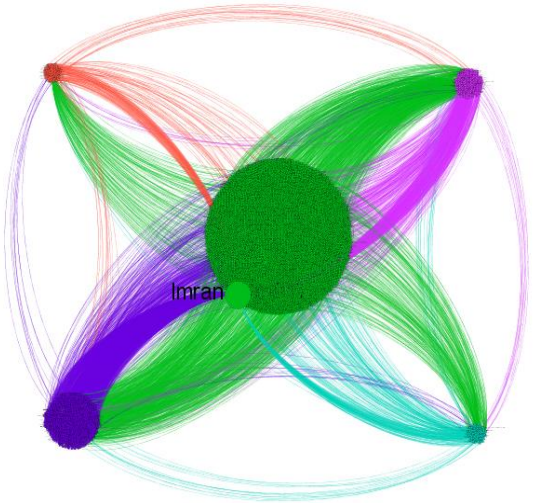
Bilawal Bhutto



74.64% of the total
50.08% = BB, 7.47% = media cell
6.72%,3.53%,3.02% and 2.90% = Party Members,
3.82% = MN, 3.02% = IK

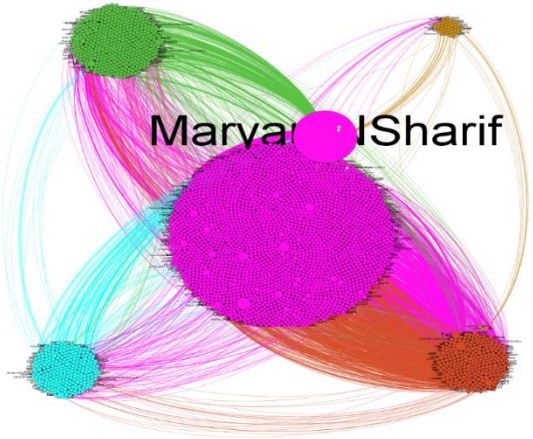
RETWEET NETWORK

Imran Khan



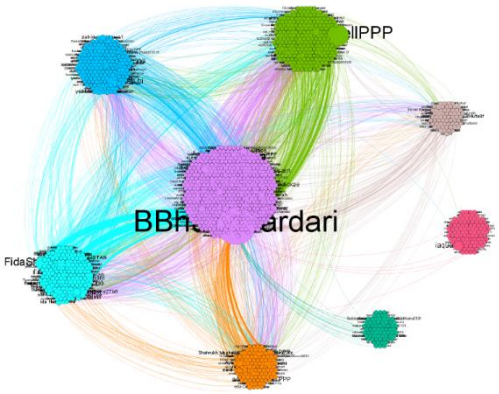
94.90% of the total network
80.16% = IK, 10.43% =official page PTI,
2.95%, 1.23%, 1.13% = other members

Maryam Nawaz



89.35% of the total network
64.46% = MN, 11.02% = media cell PMLN
6.5%, 6.38%, 0.99% = other party members

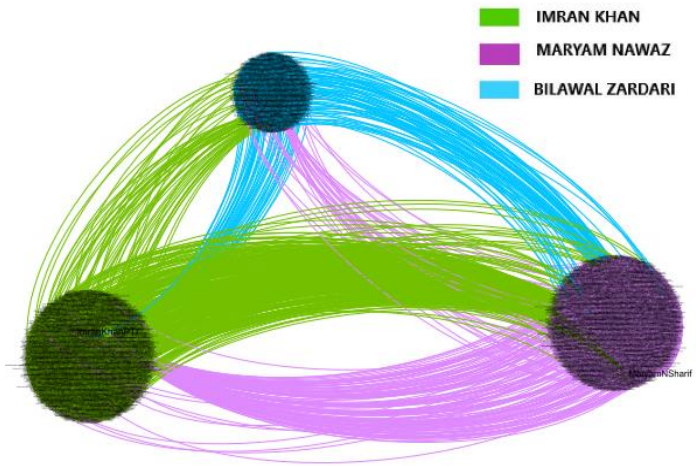
Bilawal Bhutto

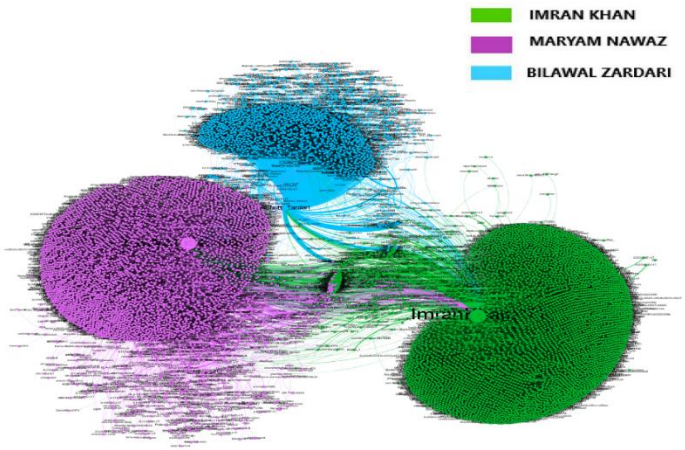


71.31% of the total network
24.90% = BB, 12.14% = media cell,
The remaining other members

COMMUNICATION LINKS (REPLY)

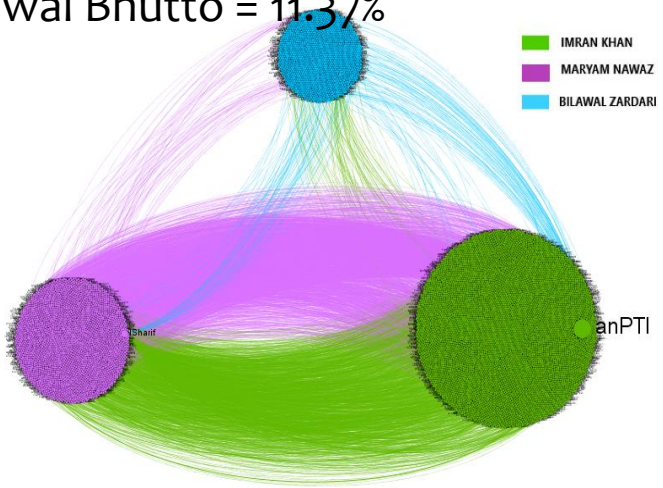
81.79% of the total network
Imran Khan = 34.70%
Maryam Nawaz = 34.52%
Bilawal Bhutto = 12.47%

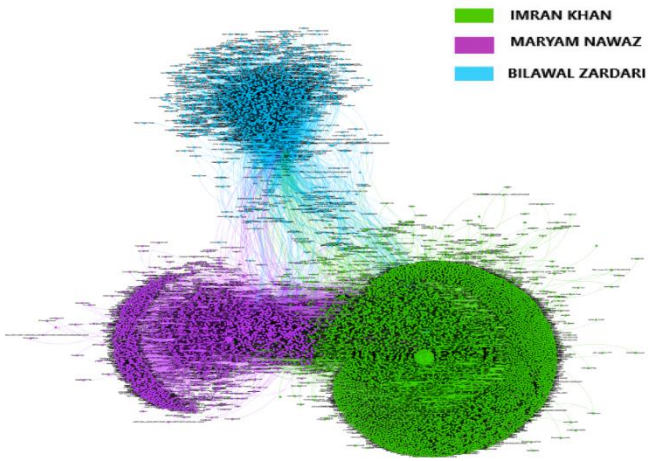




COMMUNICATION LINKS (RETWEET)

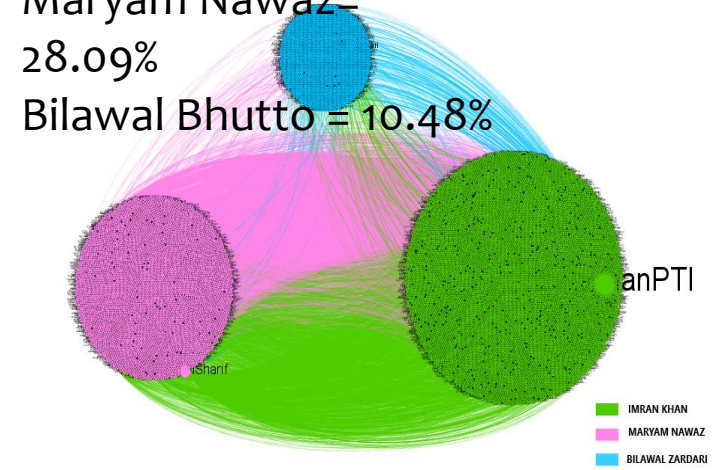
82.83% of the total network
Imran Khan = 51.26%
Maryam Nawaz= 20.20%
Bilawal Bhutto = 11.37%

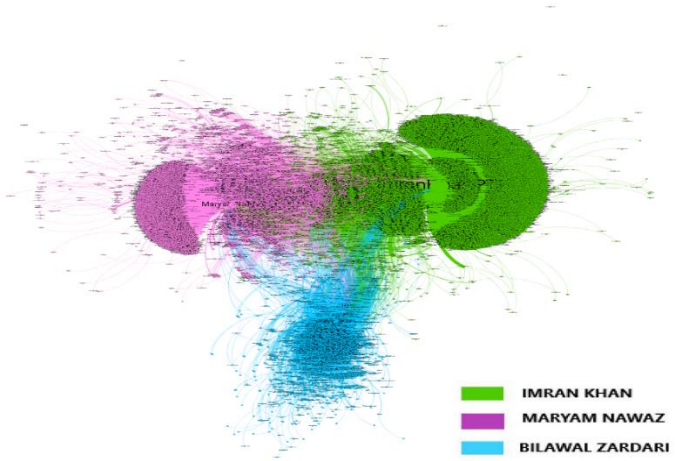




COMMUNICATION LINKS (ENTIRE NETWORK)

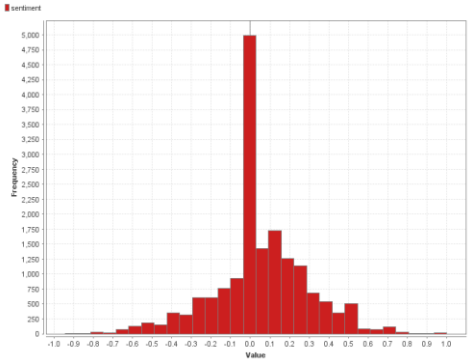
90.87% of the total network
Imran Khan= 52.30%
Maryam Nawaz= 28.09%
Bilawal Bhutto = 10.48%





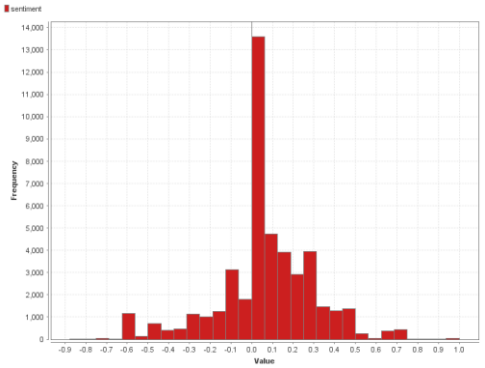
LEXICAL ANALYSIS

Imran Khan



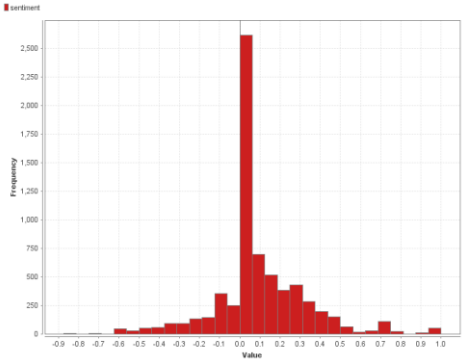
Avg sentiment = 0.052
Deviation = 0.238

Maryam Nawaz



Avg sentiment = 0.058
Deviation = 0.241

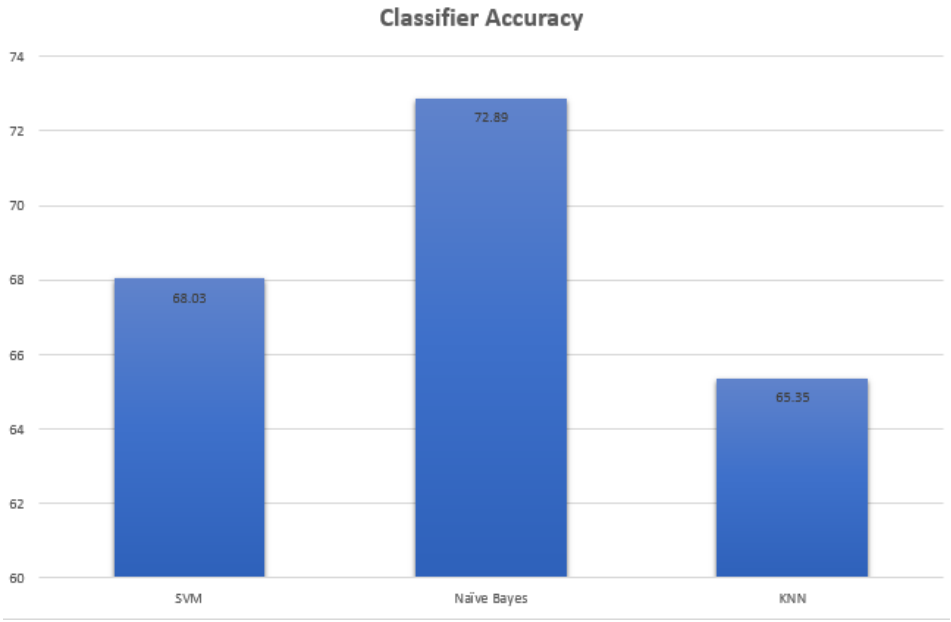
Bilawal Bhutto



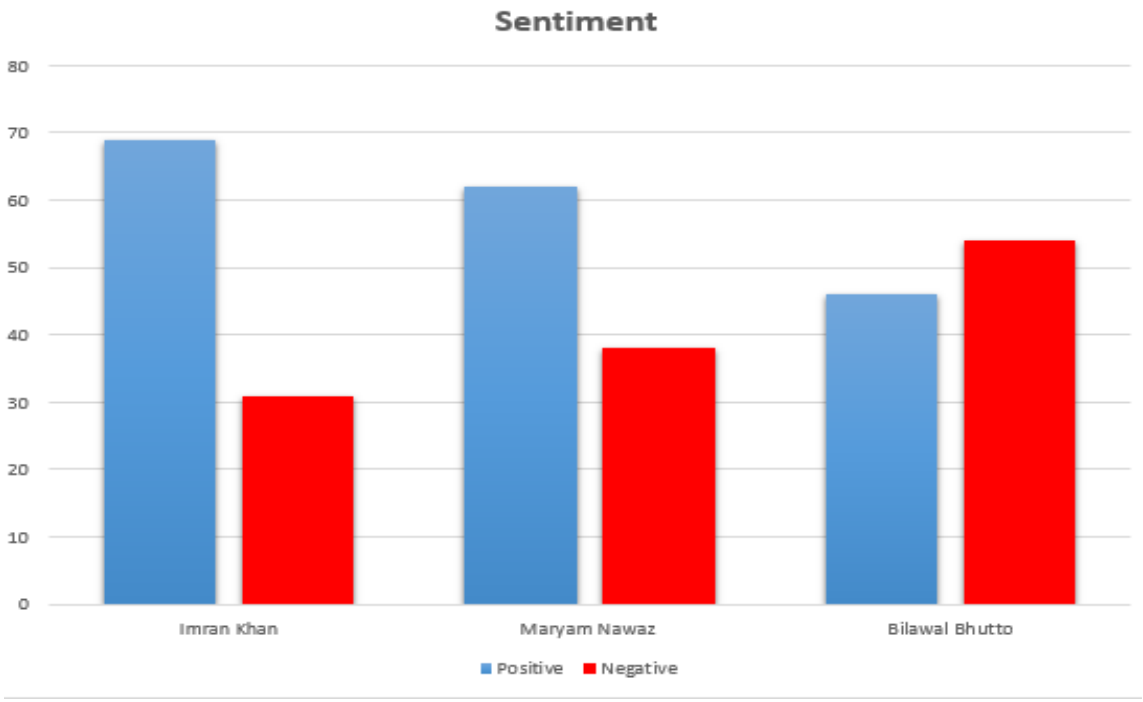
Avg sentiment = 0.083
Deviation = 0.229

MACHINE LEARNING

- 16000 Tweets Manually labeled
- Naïve Bayesian, KNN, and SVM were tested



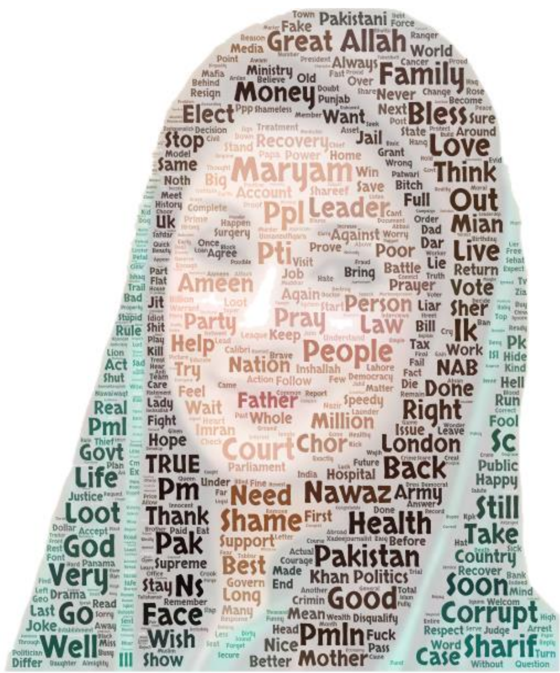
MACHINE LEARNING ANALYSIS
(Prediction)



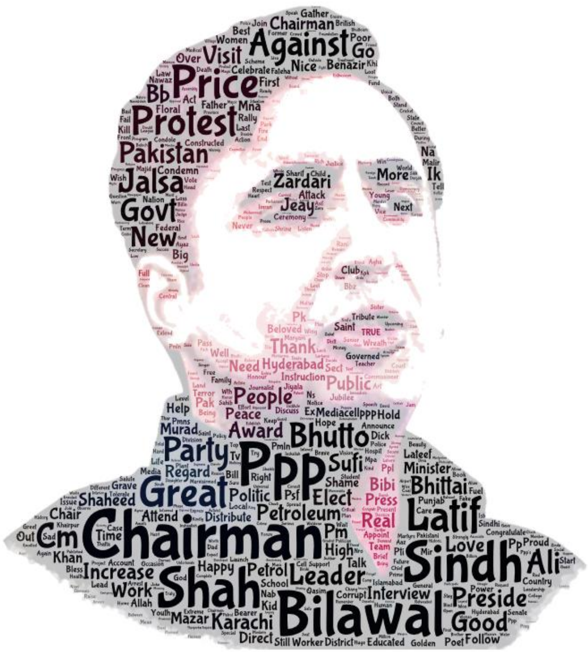
WORD CLOUD

Imran Khan

Maryam Nawaz



Bilawal Bhutto



CONCLUSION

- This study successfully identified communities
- Predicted Strongest Supporter Community
- Predicted sentiments of politician's network
- Imran Khan maybe the next Prime Minister
  - Strongest Supporter Network
  - Flexible behavior of Supporters
  - High rate of Positive Sentiment

CONCLUSION

- Politicians Behaviors
  - Imran Khan uses Twitter precisely
  - Maryam Nawaz has an excessive use of Twitter
- Each party's media cell and other party members support these politicians, especially Bilawal Bhutto
- Some tweets falling in English category have some Urdu words, that effect the performance
- Fake supporters
- People of Pakistan are flexible and are not isolated. Communication links exists

感谢关注聆听！

张华平
Email: kevinzhang@bit.edu.cn
微博：@ICTCLAS张华平博士
实验室官网：
http://www.nlpir.org

大数据千人会