社交网络舆情源发现与分析

Group Detection and Analysis in New Social Media

张华平 博士 副教授

大数据搜索与挖掘实验室 主任

kevinzhang@bit.edu.cn

@ICTCLAS张华平博士

2019.10

北京理工大學

BEIJING INSTITUTE OF TECHNOLOGY

相关科研课题支持

- 面向互联网信息的司法舆情监测与分级预警技术研究及系统研发 2018YFC0832304 “公共安全风险防控与应急技术装备”重点专项（司法专题任务）
- 语义主题与社交关系融合的特定群体发现关键技术研究 61772075 国家自然科学基金面上项目 2018.1-2021.12
- 社交网络分析及信息传播理论在舆情预警方面的示范验证 2013CB329606 国家973重点基础研究发展计划
- 基于主体个性化的微博情感分析关键技术研究 61272362 国家自然科学基金面上项目 84

纲要

舆情源分析

I 舆情困局变革：从“舆情”到“舆情源”

II 社交网络舆情源发现与分析

III 舆情源分析实战案例

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

BBS 1995

IM 1999

博客 2000

视频

社区 2005

微博 2009

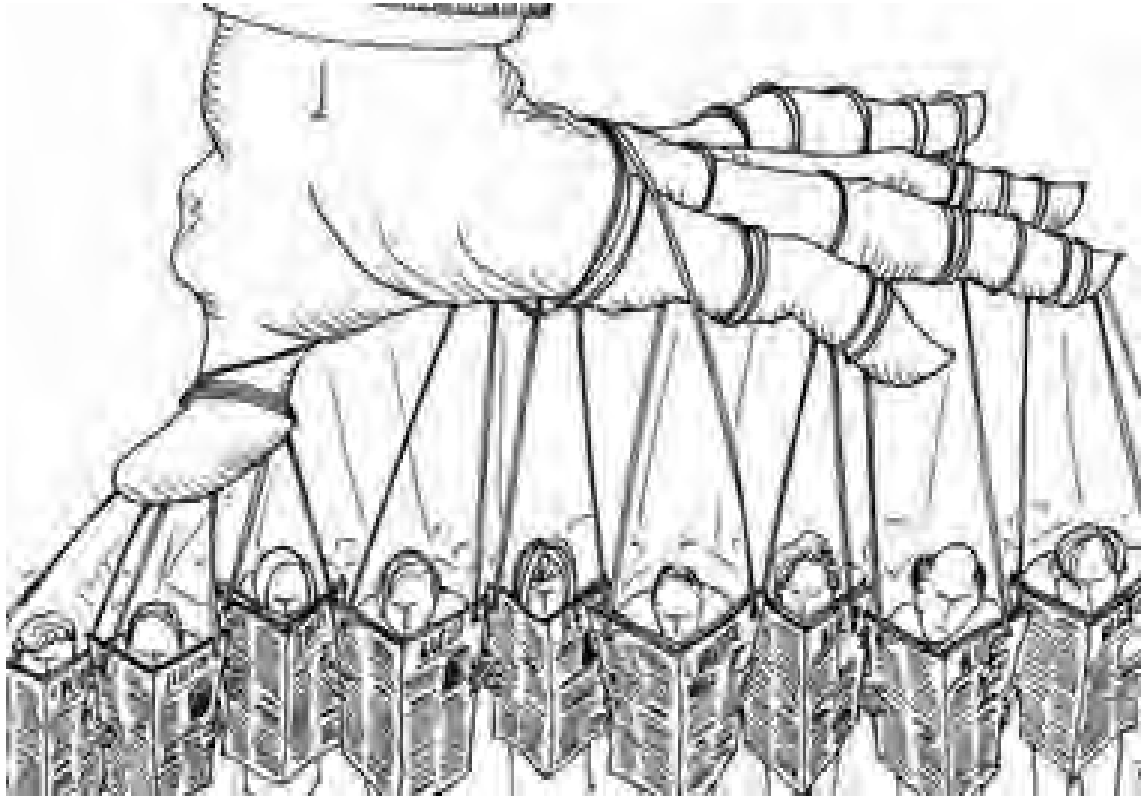
微信 2011

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

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

輿情源操控–剑桥分析的背后



輿情源操控

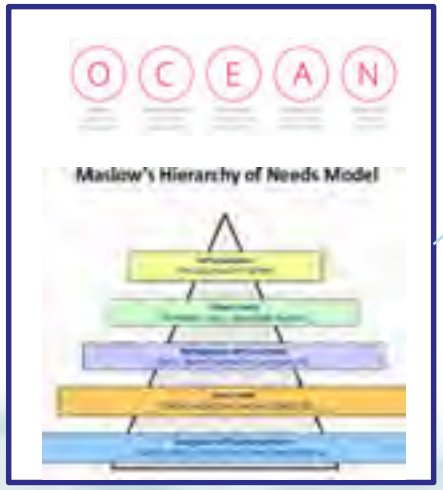
Cambridge Analytica

RUSSIAN INSTITUTE FOR STRATEGIC STUDIES

ISIS

Opinion Manipulators

Psychological Models



Targeted Community

Social Media

Computational Models

Public Opinion Manipulated

北京理工大學
BEIJING INSTITUTE OF TECHNOLOGY

剑桥分析：舆情源心理洞察

Conversations that move people

When you go beneath the surface and learn what people really care about you can create fully integrated engagement strategies that connect with every person at the individual level.

Same demographics, different personalities

Female
25-35 Years old
AMEX User

Openness
Conscientiousness
Extraversion
Agreeableness
Neuroticism

People with high openness and extraversion love new experiences they can share with lots of people.

Female
25-35 Years old
AMEX User

Openness
Conscientiousness
Extraversion
Agreeableness
Neuroticism

People with low openness and extraversion really value down time spent with their closest friends.

We Call This Behavioral Microtargeting

Discover. Understand. Engage. Repeat.

Combine our full suite of data-driven audience insight and engagement techniques with our unique and powerful Behavioral Microtargeting service that constantly learns, improves and delivers.

With Behavioral Microtargeting you'll be able to anticipate the needs of your customers and predict how their behavior will change over time, so you can build services, products and campaigns they really love.

Geographic View

Demographic View

Psychographic View

北京理工大學
BEIJING INSTITUTE OF TECHNOLOGY

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）

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

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

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？

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**
  - ⑩国家安全危害分子：台独、藏独、疆独、港独、民运、邪教、暴恐；
  - ⑩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；
  - ⑩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**
  - ⑩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）；
  - ⑩愤青；
  - ⑩理性质疑者
  - ⑩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

纲要

舆情源分析

I 舆情困局变革：从“舆情”到“舆情源”

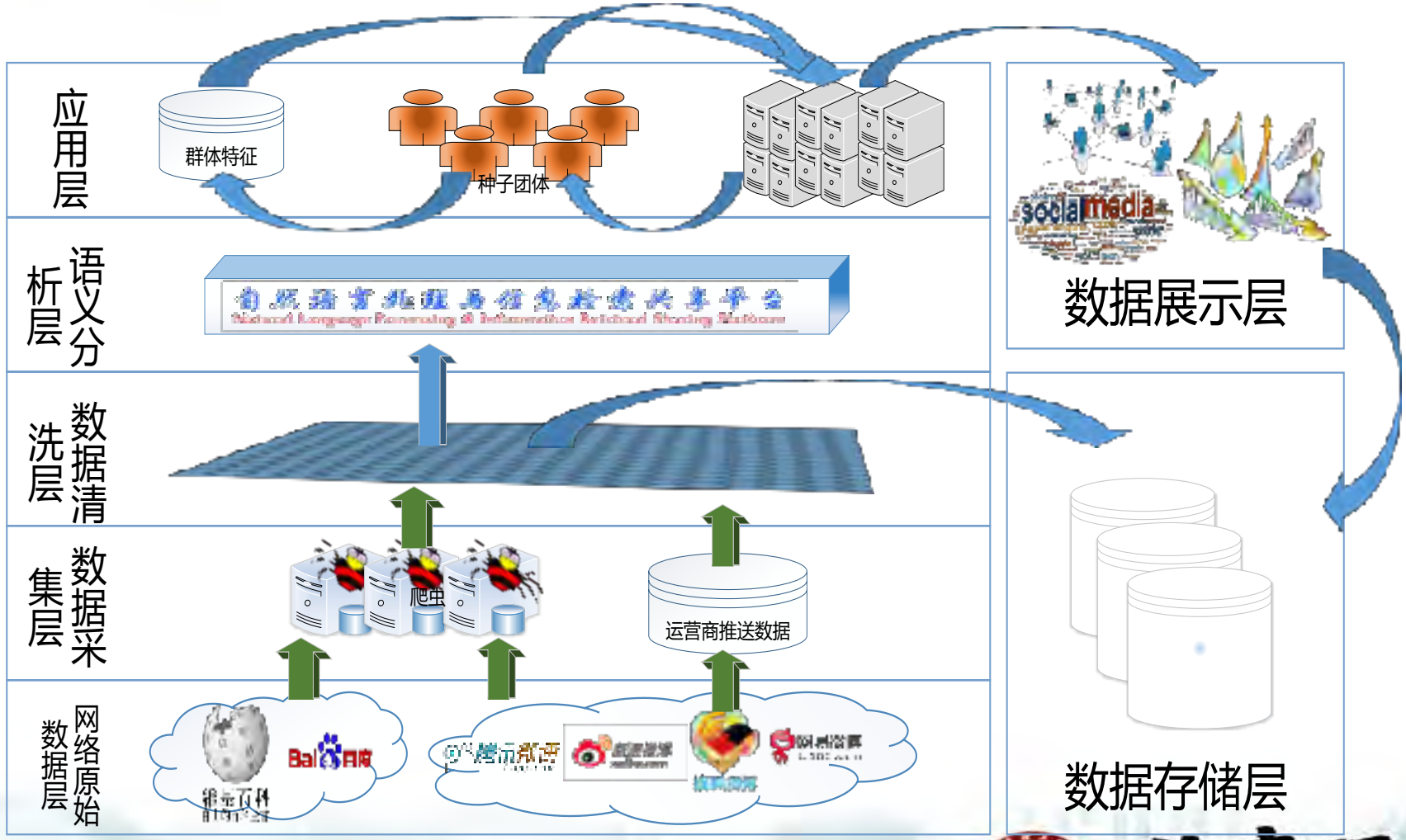
II 社交网络舆情源发现与分析

III 舆情源分析实战案例

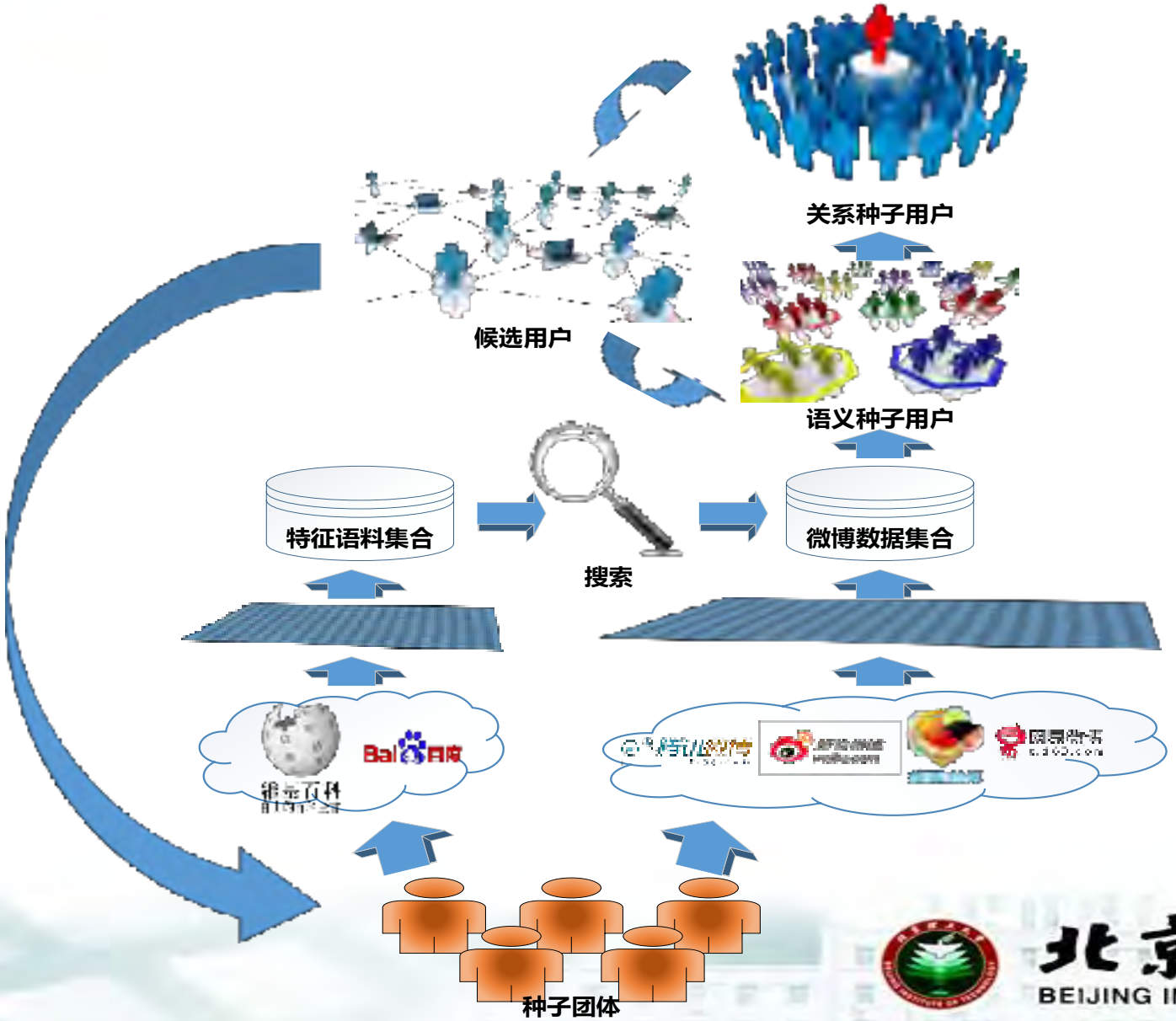
社交网络舆情源发现与分析

- 舆情源发现
  - 实现特定群体社交网络账号的发现；
- 舆情源分析
  - 实现对特定群体的社交属性、活动属性、位置属性等的全特征计算；对特定小众化群体进行快速搜索、关联分析和属性标注；
- 舆情与源关联分析
  - 特定舆情事件参与人员关联分析，舆情源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；
- 基于舆情源追踪的预警
  - 针对已构建的特定群体，初步实现对突发事件的预警预报及发展态势研判；

舆情源发现架构


应用层
群体特征
种子团体
语义分析层
数据清洗层
数据采集层
爬虫
运营商推送数据
网络原始数据层
数据展示层
数据存储层

基于语义与关系的舆情源发现

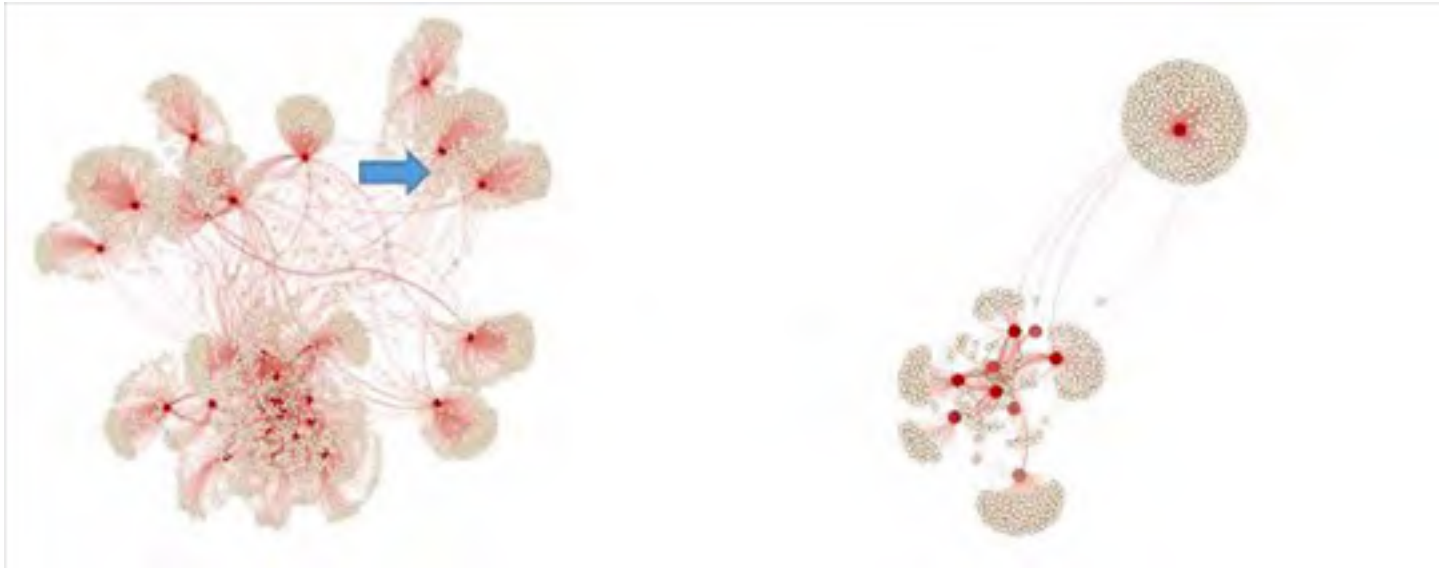


舆情源发现：失独老人发现

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。

舆情源发现：失独老人算法迭代演变过程



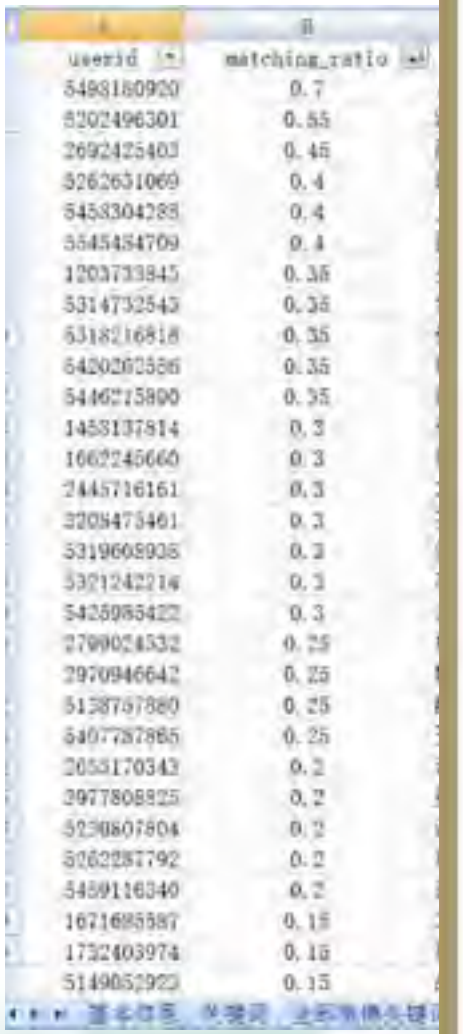
语义种子用户的关系网络的迭代演变过程

	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

舆情源发现：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

舆情源发现：犯罪网络“转世党”去哪了？



userid	matching_ratio
5488160920	0.7
5202496301	0.55
2692425403	0.45
5262631069	0.4
5458304288	0.4
5545434709	0.4
1203733845	0.35
5314732543	0.35
5318216818	0.35
5420202586	0.35
5446275890	0.35
1453137814	0.3
1662245660	0.3
2445716161	0.3
3208473461	0.3
5319609936	0.3
5321242214	0.3
5426965422	0.3
2799024532	0.25
2970946642	0.25
5138757880	0.25
5407787865	0.25
2053170343	0.2
2977808825	0.2
5230807804	0.2
5262287792	0.2
5459116340	0.2
1671685587	0.15
1732403974	0.15
5149052923	0.15

北京理工大學
BEIJING INSTITUTE OF TECHNOLOGY

舆情源分析：博主行为建模

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	0	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	1	3	0	1	2	2	1	0	0	23
.....																									
2011/10/8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	2
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
LOC2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

舆情源分析：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	0.843321876	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	0.768093971	[illegible]	0.797002832	0.641049731
算术平均差异率	0.692160456	0.770269605	0.769741208	0.771522319	0.811455006	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%

舆情源分析：博主行为分析







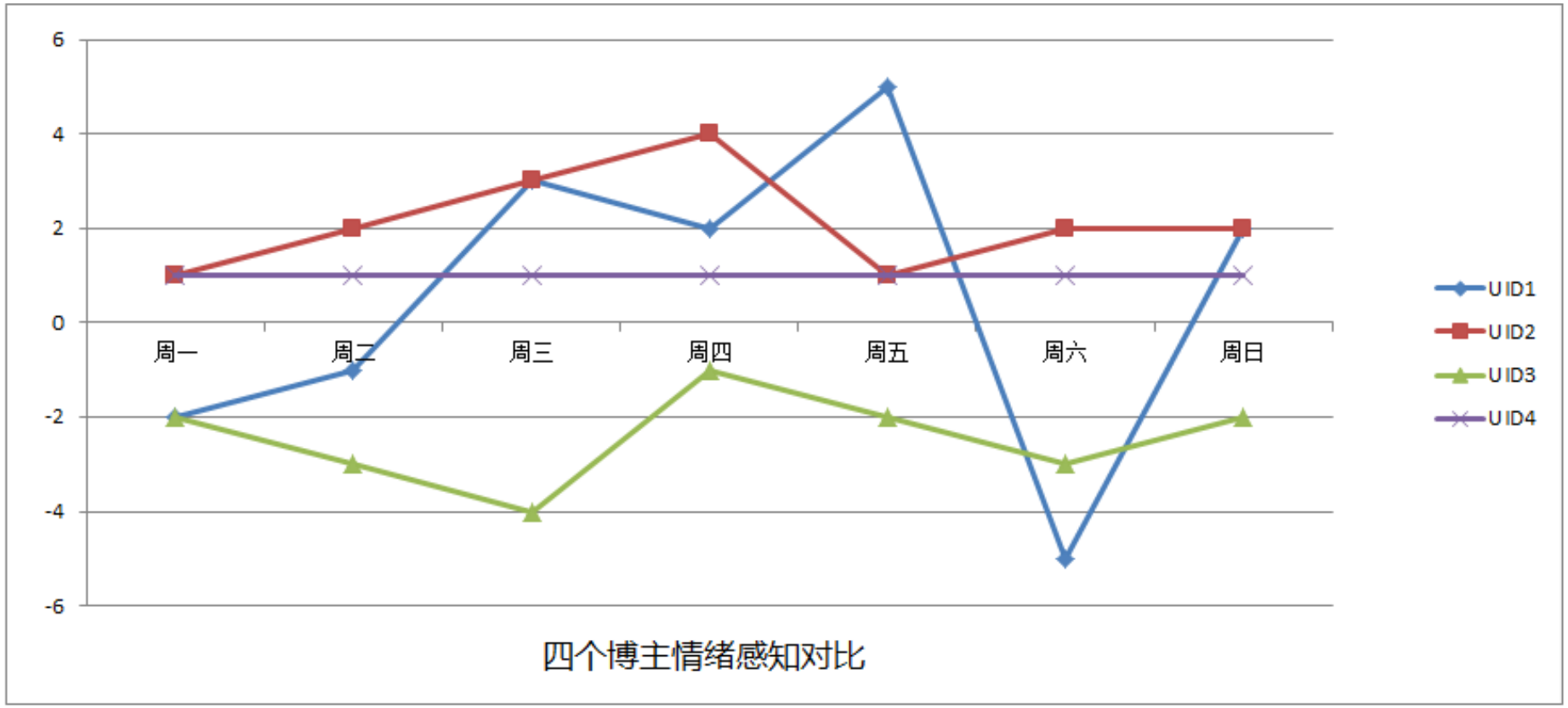


	张华平	任志强	潘石屹	张鸣	白硕	林伯强	张栋	方文山	刘强东
张华平	1								
任志强	0.447339	1							
潘石屹	0.746915	0.760761	1						
张鸣	0.84744	0.612968	0.818428	1					
白硕	0.698806	0.644066	0.81019	0.704533	1				
林伯强	0.603462	0.343865	0.602498	0.813252	0.482863	1			
张栋	0.773073	0.465831	0.758745	0.765128	0.843614	0.700826	1		
方文山	0.073967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

舆情源分析 :微博博主价值观分析(Big 5模型)

id	昵称	身份	pid	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	退从向量
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	====	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	====	1	10	3	2	8	6	5	9	4	7
65	ICTCLAS张华平博士	张华平博士的	====	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	====	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	====	2	10	1	7	8	9	5	4	3	6
95	我不叫成书华	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	====	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	====	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	====	1	9	4	2	8	6	10	5	3	7

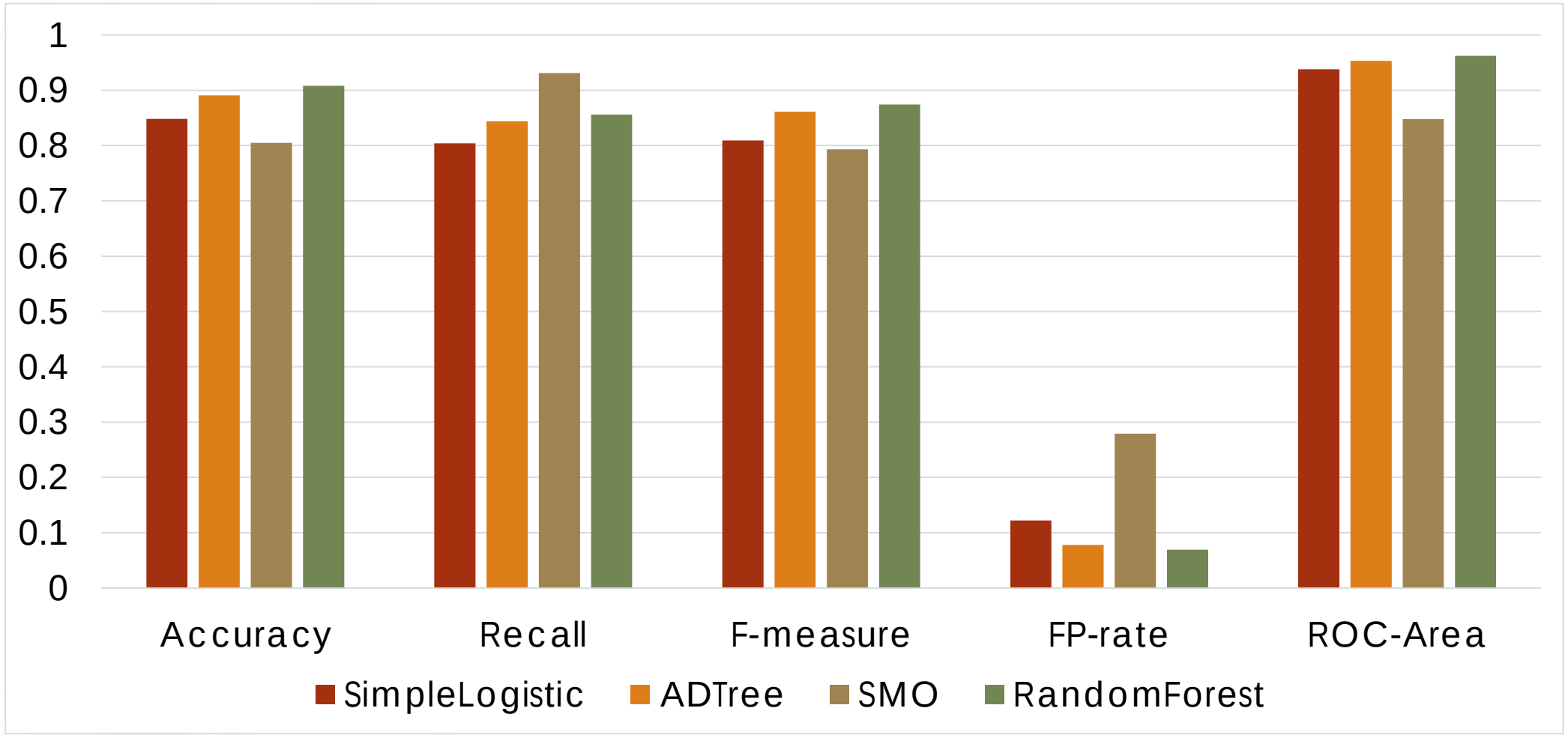
舆情源分析：微博博主情绪感知



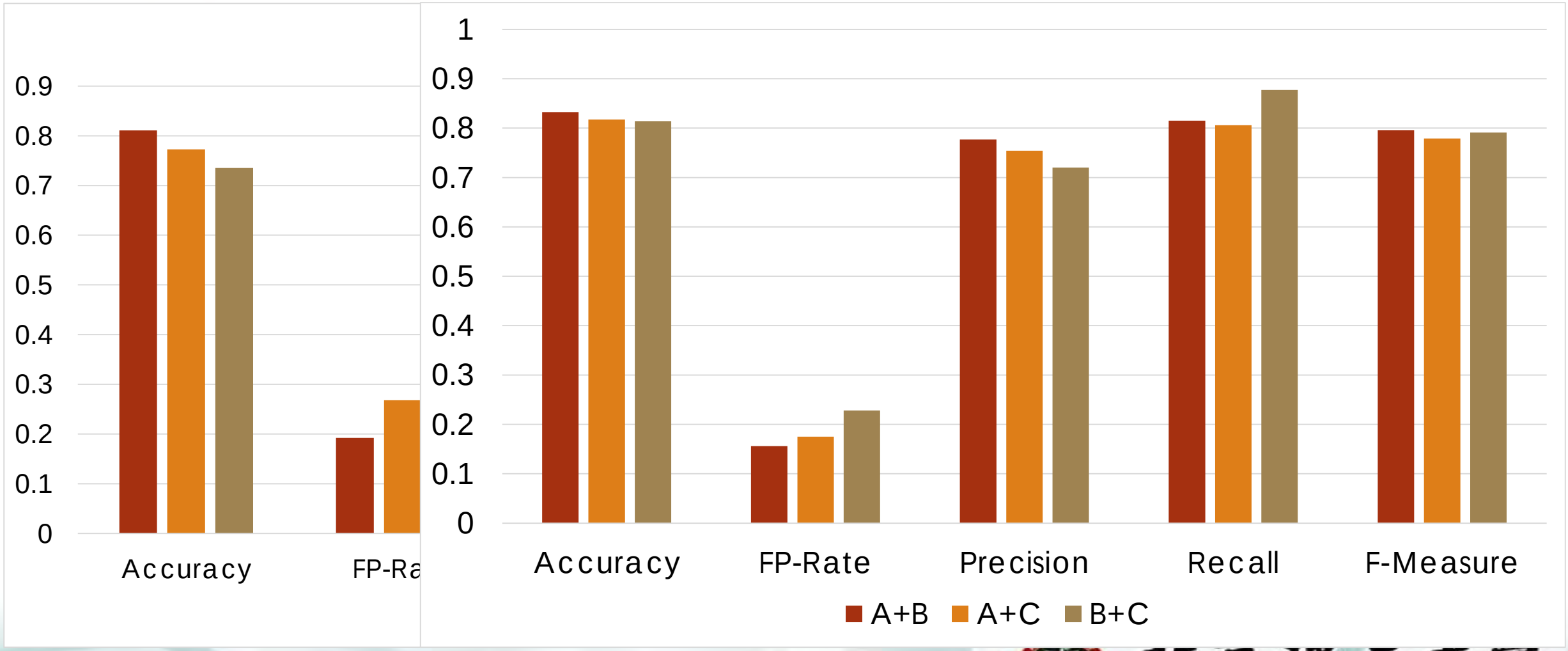
舆情源分析：水军检测

行为特征	内容特征	属性特征
发帖频率(PF)	平均“@”数(MFF)	粉丝数(FAN)
上网方式(TPWN)	转发微博比例(RMR)	关注数(FON)
转发上网方式(TRWN)	原创微博比例(OMR)	粉丝关注比(FDF)
参与大于100转发(PI100)		

舆情源分析：水军检测


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Accuracy
Recall
F-measure
FP-rate
ROC-Area
SimpleLogistic
ADTree
SMO
RandomForest

舆情源分析：水军检测


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Accuracy
FP-Rate
Precision
Recall
F-Measure
A+B
A+C
B+C

舆情源分析：水军检测

	Without our Features			All Features		
Classifier	Accuracy	FP-Rate	F-Measure	Accuracy	FP-Rate	F-Measure
Simple Logistic	73.75%	0.356	0.728	84.83%	0.122	0.809
AD Tree	83.66%	0.116	0.834	89.08%	0.078	0.861
SMO	63.08%	0.161	0.409	80.50%	0.279	0.793
Random Forest	85.5%	0.095	0.843	90.08%	0.069	0.874

舆情源分析：谣言检测

分类器	Accuracy	Recall	ROC Area
Simple Logistic	58.67%	0.554	0.579
AD Tree	62.00%	0.594	0.611
SMO	55.45%	0.620	0.556
Random Forest	65.44%	0.633	0.670

纲要

舆情源分析

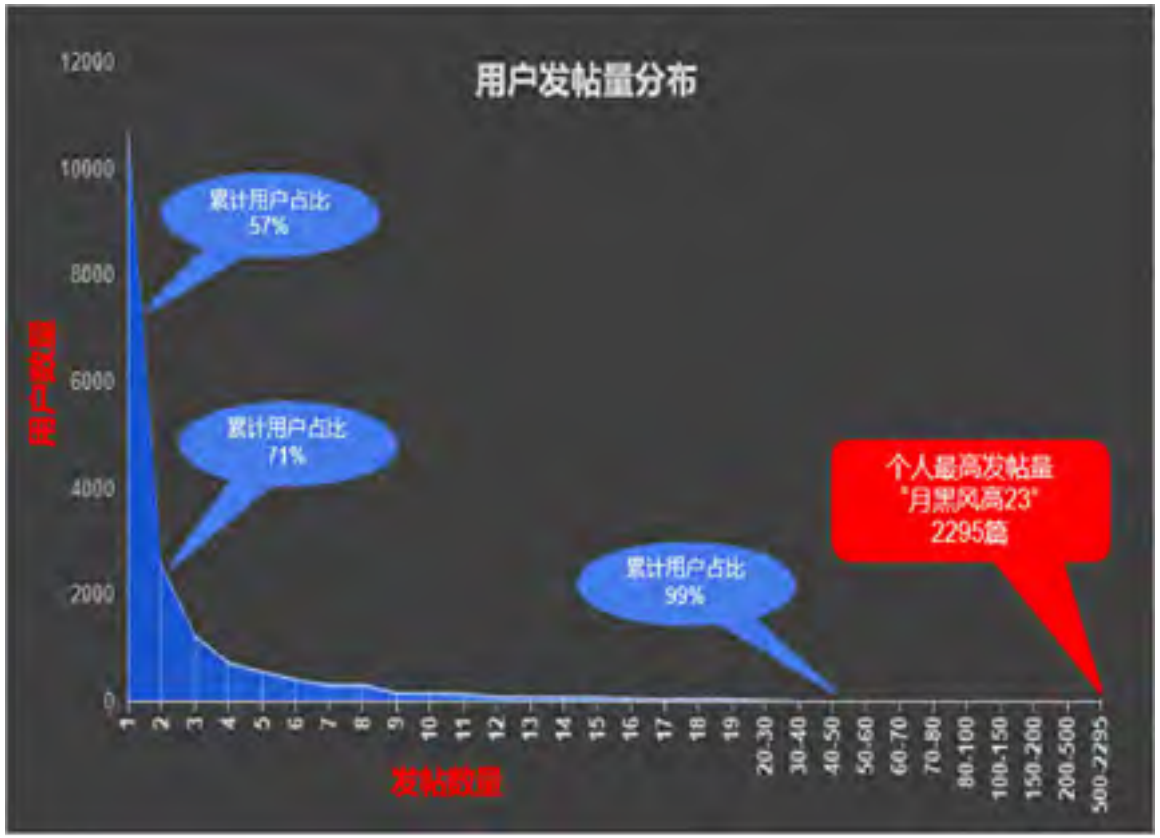
I 舆情困局变革：从“舆情”到“舆情源”

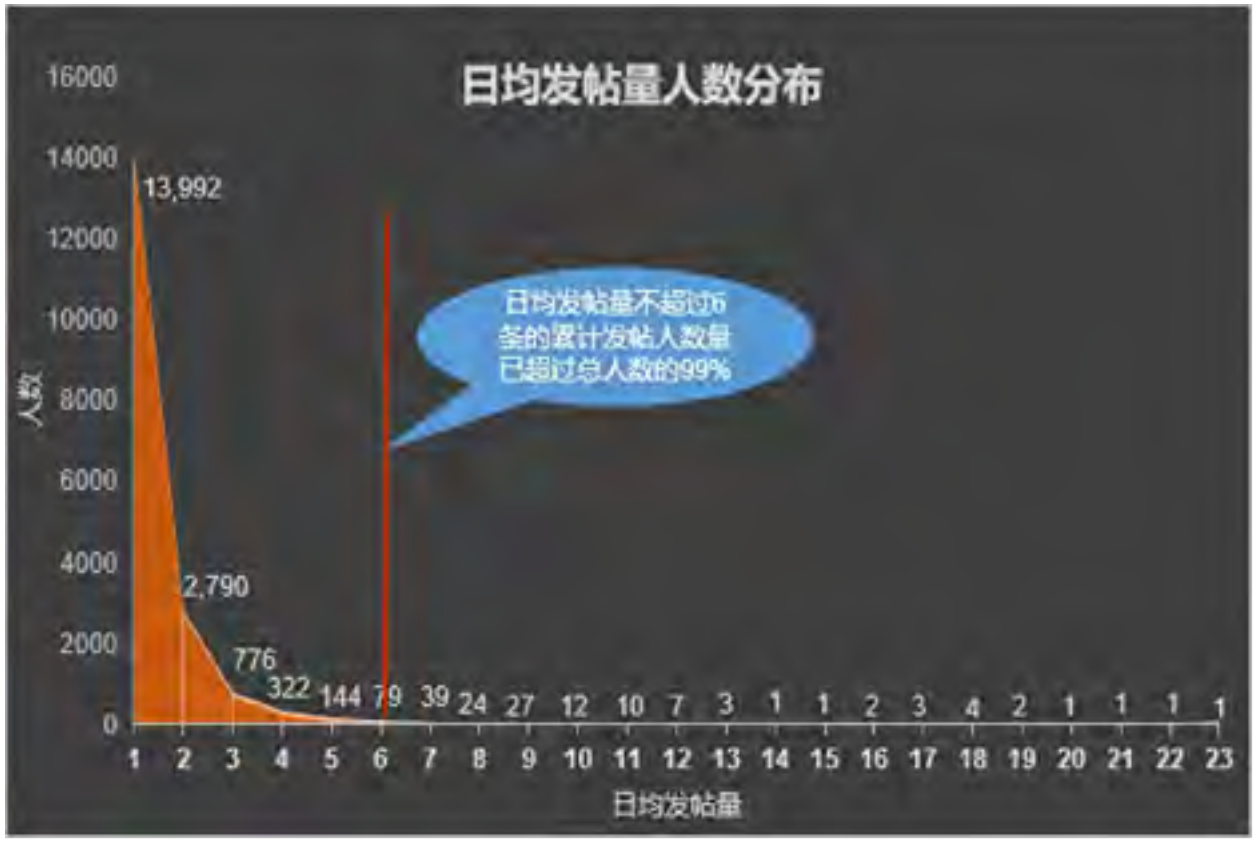
II 社交网络舆情源发现与分析

III 舆情源分析实战案例

某汽车品牌舆论场画像

发帖人总量：**18645个**；发帖总量：**87484条**；人均发帖量：**4.7条/人**





北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

某汽车品牌舆论场画像：从2万到200

发帖总量与日均发帖量都异常

发帖总量异常账号　　日均发帖量异常账号

某汽车品牌舆论场画像：极端黑公关

发帖人“追逐人生LXB”语义画像

敌对攻击号

659 篇

30 条

-205

个

发帖人“长安非常不安”语义画像

601 条

每天 50 条

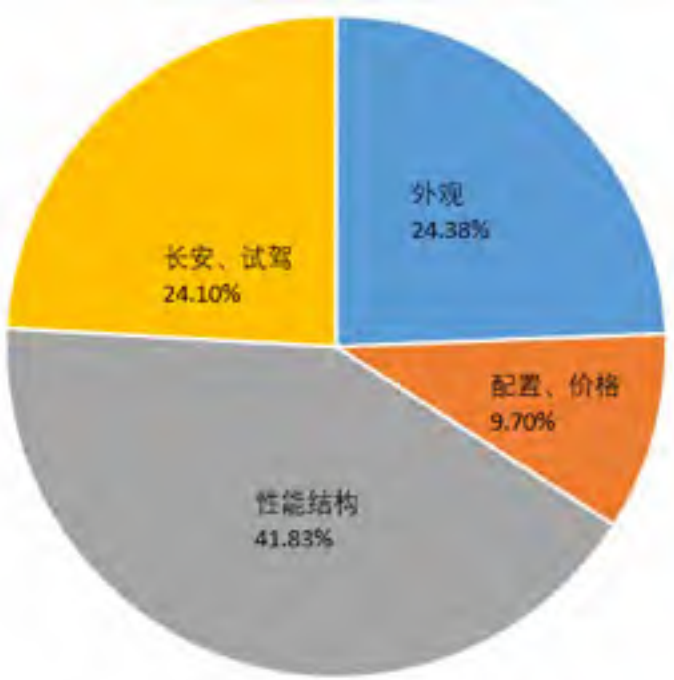
-710

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

原贴语义分析

通过对原贴语义分析，可以得出用户的关注热点。

原贴关键词



关注热点分布

性能结构的词语在发帖中占比最高，达到48.83%，其次是外观的词语，占比24.38%，长安试驾词占比24.10%，而配置、价格，占比为9.70%。因此，最受关注的是性能结构，尤其是轮毂、后备箱、溜背等。

主观舆情分析

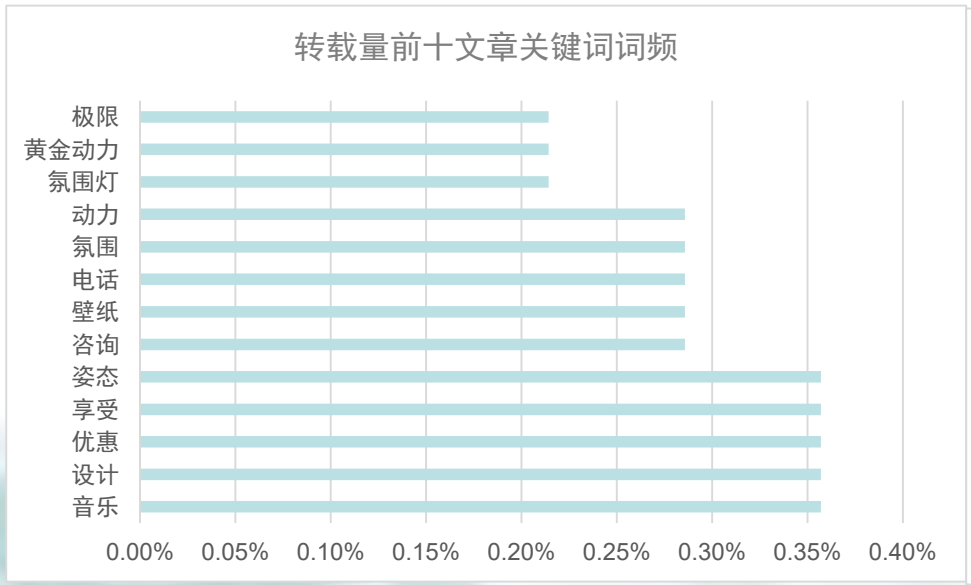
主观舆情是标题含有**的媒体舆情，对于这部分舆情，首先分析主观舆情的总体健康程度

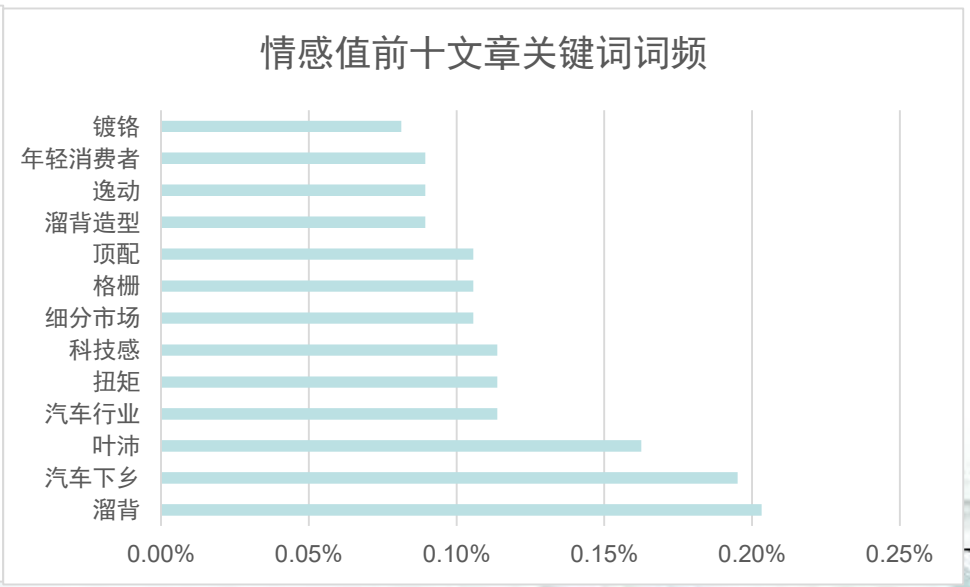
在我们拿到的主观报道的所有的11352篇文章中，平均得分27.90分，其中，正面评分文章共有9987篇，平均分为31.76分，占比87.98%，负面评分文章共有238篇，负面得分-1.79 占比2.1% ，剩下有1127篇文章情感值为0，情感为中性，占比9.93%。以下为文章得分前五关键词和网站所有文章总分前五的关键词。

通过这部分关键词可以分析情感值较高的文章的描述热点。

主观舆情分析

在文章层面，对传播量较大的文章和情感得分高的文章分别进行分析：

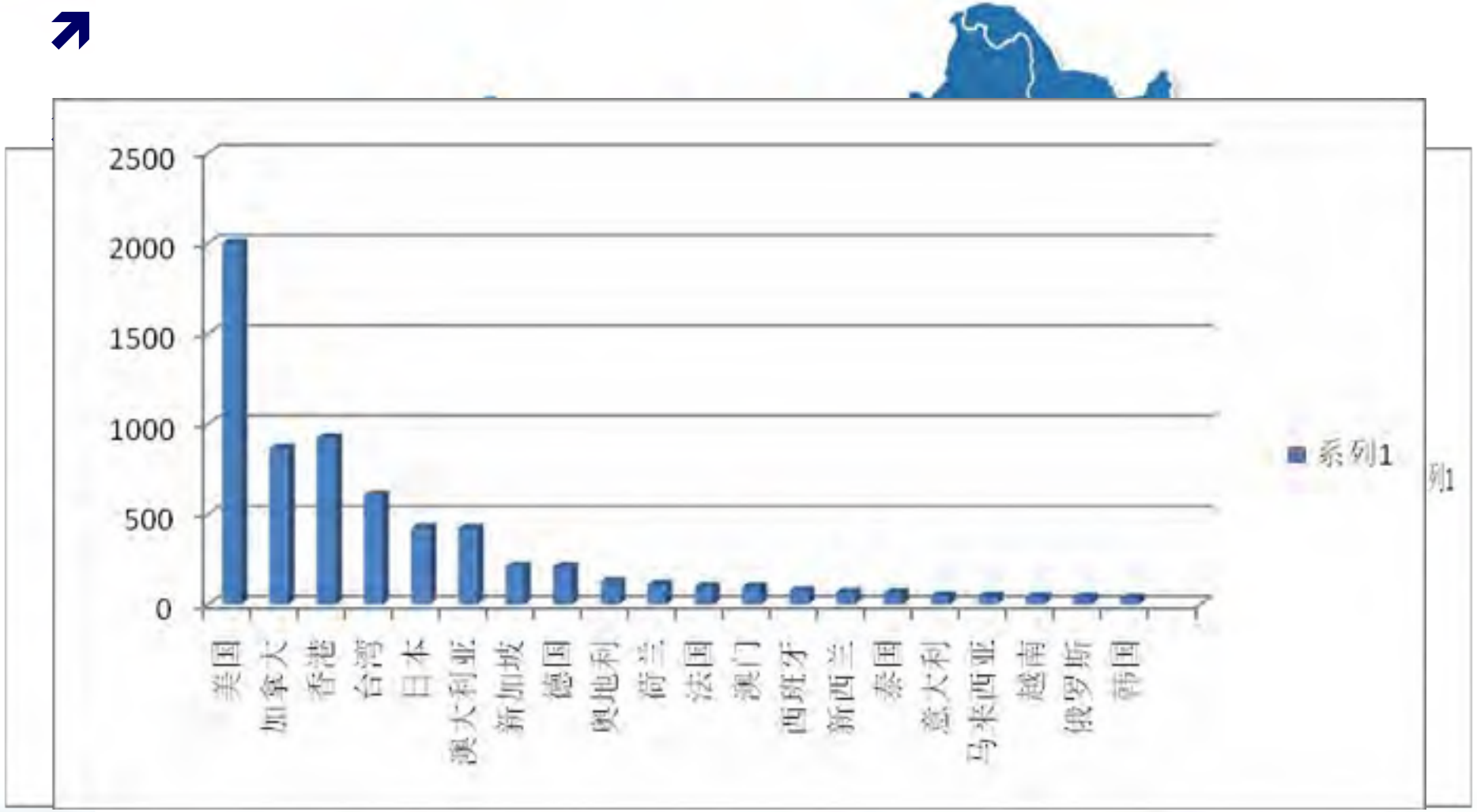




情感值得分较高的帖子中关于车的配置的词提及的较多，而在转载量高的文章中，关于车的性价比、服务、试驾体验的词提及较多

“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



所有参与者的观点分析

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

草根的观点分析

大V的观点分析

媒体的观点分析

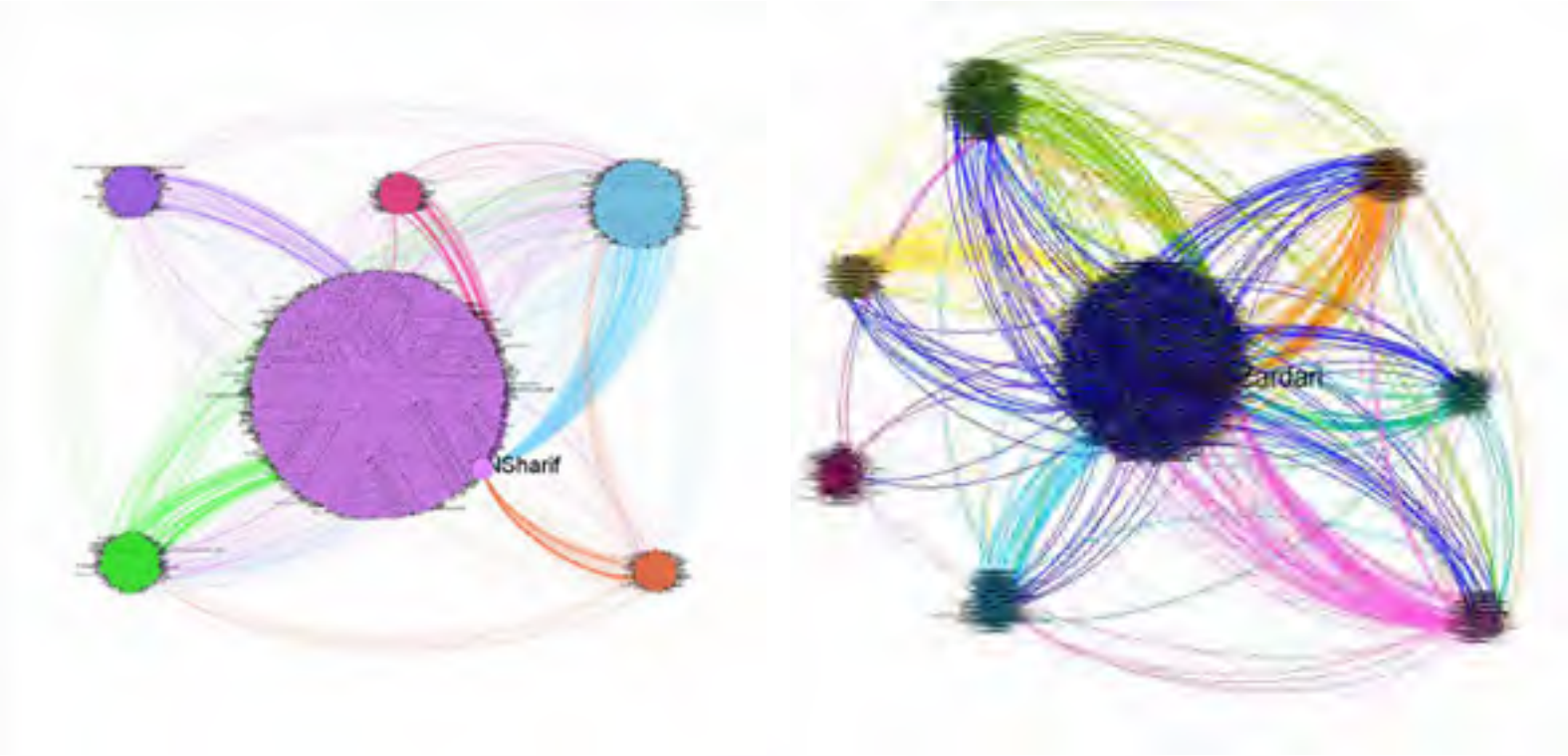
常胜将军
抗战时期
链接
将军
张灵甫
微博
将军
英雄
王耀武
没有
现在
张道宇
网易新闻
剿匪
历史人物
国民党
名将
战功
评价
回复
山东媒体
看来
解放军
抗日将领
少将
照片
文章
知道
谢谢
先生
人物
军人
抗日
抗战
Copyright 2015
tagxedo.com

SNA Research: Pakistan Politics Leader

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

SNA Research: Pakistan Politics Leader

Imran Khan

Maryam Nawaz

Bilawal Bhutto

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

感谢您的耐心聆听！

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

北京理工大學
BEIJING INSTITUTE OF TECHNOLOGY