



北京交通大学  
BEIJING JIAOTONG UNIVERSITY



# 《大数据概论》 大数据行业应用

鲍鹏  
软件学院





# 大数据与社交媒体的融合

- 随着社会网络服务的发展，社交媒体作为人们传播信息和表达观点的重要渠道，包含大量丰富的有用信息，这些信息伴随社交媒体服务的兴起，形成了各种各样的社交媒体数据，比如微博类网站的文本信息流数据、媒体分享网站的多媒体数据、社交网站的用户交互数据、签到网站的地理位置数据、购物网站的消费数据等等，这些社交媒体数据已成为大数据最具代表性的数据来源之一。这些社交媒体多源数据从不同角度记录着人们的网络生活,并映射着物理世界。



# 大数据与社交媒体的融合

- 什么是社交媒体
- 社交媒体大数据的分析与挖掘
- 社交媒体大数据的未来挑战



# 大数据与社交媒体的融合

- 什么是社交媒体





# 社交媒体的定义

## 中国十大网络社交平台-社交媒体平台-移动社交平台，最受欢迎的网络社交应用<2018>

1	微信	( 0755-86013388 , 时下最热门的聊天通讯软件,腾讯移动互联网应用领域的看家产品,主打熟人圈的社交媒体,其朋友圈/微...	↑↑ ↓↑
2	QQ	( 0755-83765566 , 中国大陆即时通讯市场的王者,红色围巾的小企鹅为其典型标志,手机用户基本装有的交友软件,深圳市腾	↑↑ ↑↑
3	微博	( 400-6900000 , 曾用名新浪微博,国内较大的娱乐休闲生活服务信息分享和交流平台,媒体监控和跟踪突发消息的重要来源,	↑↑ ↑↑
4	QQ空间	( 0755-86013388 , QQ用户的网上家园,展现个人特色的多媒体空间博客,活跃度高/互动性强的生活记录平台,人们常用的大	↑↑ ↑↑
5	百度贴吧	( 010-59928888 , 百度旗下独立品牌,全球较大中文社区,基于关键词的主题交流社区,众多网络流行用语的发源地,北京百度	↑↑ ↑↑
6	知乎	( 010-61190680 , 中文互联网高质量内容社区,国内知名网络问答社区,以高质量多样性著称,高成长社交媒体平台,北京智者	↑↑ ↑↑
7	抖音短视频	( 010-58732757 , 今日头条旗下人气音乐创意短视频社交软件,专注年轻人的15秒音乐短视频社区,国内领先的短视频APP,...	↑↑ ↑↑
8	豆瓣douban	( 文艺青年聚集的社交平台,以影评书评和快速更新的影音资讯而著称,集知识性和互动性为一体的文化社交媒体,北京豆网科	↑↑ ↑↑
9	MOMO陌陌	( 028-62836666 , 基于地理位置的开放式移动视频社交应用,广泛交友的好工具,可以通过视频/文字/语音/图片来展示自己,...	↑↑ ↑↑
10	探探	( 基于大数据智能推荐/全新互动模式的社交App,较受年轻人欢迎的社交应用,主打陌生人社交,2018年被陌陌收购,探探文化...	↑↑ ↑↑



# 社交媒体的定义

- 微信，时下最热门的聊天通讯软件，腾讯移动互联网应用的看家产品，主打熟人圈的社交媒体，其朋友圈/微信红包/公众号等成为人们日常生活的焦点



- 微信是腾讯公司于2011年1月21日推出的一个为智能终端提供即时通讯服务的免费应用程序，微信支持跨通信运营商、跨操作系统平台通过网络发送免费（需消耗少量网络流量）语音短信、视频、图片和文字，同时，也可以使用通过共享流媒体内容的资料和基于位置的社交插件“摇一摇”、“漂流瓶”、“朋友圈”、“公众平台”、“语音记事本”等服务插件。
- 微信提供公众平台、朋友圈、消息推送等功能，用户可以通过“摇一摇”、“搜索号码”、“附近的人”、扫二维码方式添加好友和关注公众平台，同时微信将内容分享给好友以及将用户看到的精彩内容分享到微信朋友圈。





# 社交媒体的定义

- QQ，民间昵称“企鹅”，中国大陆即时通讯市场的王者，国人维系人脉的必备软件，手机用户基本装有的交友软件，红色围巾的小企鹅为其典型标志
- 腾讯QQ是深圳市腾讯计算机系统有限公司开发的一款基于Internet的即时通信（IM）软件。腾讯QQ支持在线聊天、视频电话、点对点断点续传文件、共享文件、网络硬盘、自定义面板、QQ邮箱等多种功能。并可与移动通讯终端等多种通讯方式相连。您可以使用QQ方便、实用、高效的和朋友联系，而这一切都是免费的。





# 社交媒体的定义



- 社交媒体也称为社会化媒体、社会性媒体，指允许人们撰写、分享、评价、讨论、相互沟通的网站和技术



- 社交媒体是人们彼此之间用来分享意见、见解、经验和观点的工具和平台，现阶段主要包括博客、论坛、播客等等











- [illegible]



# 社交媒体的发展







表1 社交媒体的发展

时间	发展历程	社交媒体
1971年	ARPA（高级研究项目署）项目的科学家发出世界第一封电子邮件，使用“@”区分用户名与地址。1987年9月20日中国第一封电子邮件由“德国互联网之父”维纳·措恩与王运丰在北京的计算机应用技术研究成功发送到德国卡尔斯鲁厄大学。	
1980年	新闻组诞生，简单地说就是一个基于网络的计算机组合，这些计算机被称为新闻服务器，不同的用户通过一些软件可连接到新闻服务器上，阅读其他人的消息并可以参与讨论。Usenet是分布式互联网交流系统，数以千计的人在上面讨论科技、文学、音乐和体育赛事等。	
1991年	伯纳斯·李经过多年实践和改进，提议采用一个新的信息发布协议，最终成就了以“超链接”为特征的万维网——World Wide Web。	
1994年	世界上第一个个人博客：斯沃斯莫尔学院学生Justin Hall建立自己的个人站点“Justin’s Links from the Underground”，与外部网络开始互联。Justin Hall坚持更新自己的博客坚持了11年，现在被公认为“个人博客元勋”。	
1995年	Classmates.com成立，旨在帮助曾经的幼儿园同学、小学同学、初中同学、高中同学、大学同学重新取得联系。	
1996年	早期搜索引擎Ask.com上线，它允许人们用自然语言提问，而非关键词（比如：“今天上映什么电影”，而不是“10月23日电影上映”）。	



# 社交媒体的发展








表1 社交媒体的发展

时间	发展历程	社交媒体
1997年	美国在线实时交流工具也称在线即时通讯软件AIM (AOL Instant Messenger) 上线；	
1998年	在线日记社区 Open Diary 上线，它允许人们即使不懂HTML知识也可以发布公开或私密日记。更重要的是，它首次实现人们可以在别人的日志里进行评论回复。	
1999年	博客工具Blogger出现；全球科技公司之间的专利站捧红的FOSS Patent就是用Blogger建的网站。	
2000年	Jimmy Wales 和 Larry Sanger 共同成立 Wikipedia，这是全球首个开源、在线、协作而成的百科全书，由来自世界各地的志愿者合作编辑而成，整个计划总共收录了超过2,200万篇条目，而其中又以英语维基百科以超过404万篇条目的数字排名第一。	
2001年	Meetup.com 网站成立，专注于线下交友。网站的创建者是 Scott Heiferman，2001年“9·11”事件以后，他成立了 Meetup.com 是一个兴趣交友网站，鼓励人们走出各自孤立的家门，去与志趣相投者交友、聊天。	
1997年	美国在线实时交流工具也称在线即时通讯软件AIM (AOL Instant Messenger) 上线；	



# 社交媒体的发展

表1 社交媒体的发展

时间	发展历程	社交媒体
2003年	面向青少年和青年群体的MySpace上线，它再一次刷新了社交网络的成长速度：一个月注册量突破 100 万。还有WordPress，它由全球各地的几百名网友通过在线协作创建，目前在全球已经拥有数千万用户——截止2011年12月，发布一年的 WordPress 3.0 获得了 6500 万次下载。	 
2004年	Facebook成立，根据7月Facebook上市后的首份财报Facebook目前每月有9.55亿用户活跃用户（MAU），每月移动平台活跃用户数有5.43亿。	
2005年	YouTube成立，它在成立后迅速被Google相中，2006年从Google那里得到的收购价是16.5亿美元。	
2006年	Twitter成立，由于它内容限制在140字以内，迅速成为方便的交流工具和强大的自媒体平台；成立的还有 Spotify，现在是社交音乐分享型应用的典型，拥有1500万MAU和400万付费用户。	
2007年	Tumblr成立于2007年，是目前全球最大的网站，也是轻博客网站的始祖。一种介于传统博客和微博之间的全新媒体形态，既注重表达，又注重社交，而且注重个性化设置，成为当前最受年轻人欢迎的新媒体之一。	
2008年	Groupon上线，是国际上最大的团购网站，最早成立于2008年11月，以网友为经营卖点。其独特之处在于:每天只推一款折扣产品、每人每天限拍一次、折扣品一定是服务类型的、服务有地域性、线下销售团队规模远超线上团队。	



# 社交媒体的发展

表1 社交媒体的发展

时间	发展历程	社交媒体
2009年	Foursquare 上线，以“签到”（check-in）组建基于地理位置的社交网络，Foursquare 成立于纽约市，每年 4 月 16 日在纽约拥有一个独特的“4SQ 日”。	
2010年—2011年	Google最成功的产品Gmail推出微博客和沟通工具 Google Buzz 上线，但这是一个失败的产品，2011 年12 月 15 日彻底被 Google 终结。2011 年，Google Buzz 的继承者 Google+ 上线。	
2012年	Pinterest 呈现爆发式增长，在 2011 年底被 TechCrunch 评为“年度最佳创业公司”，它是目前网站史上最快达到 1000 万独立访客的网站	
2013年	腾讯微信发展速度惊人：用户数从0到1亿，历经14个月；从1亿到2亿，用了半年；从2亿到3亿，只花了大约4个月；截至2013年10月，微信全球用户数已经超过6亿。	
2014年	Vkontakte是俄罗斯及邻国的主要社交网络，2014年Pinterest功能更强大了，增加了诸如Place Pins(结合Foursquare和Mapbox的地理位置服务)和Rich Pins(提供更丰富的图片信息)，以促进Pinterest服务变现。	
2015年—2016年	八大社交媒体：微信、微博、陌陌、知乎、Facebook、Twitter、Snapchat以及Instagram 在用户增长和商业变现上进行了不断努力尝试	





# 讨论

- 你常用的社交媒体有哪些？
- 对你的生活有什么影响？







- 社交媒体大数据的分析与挖掘





# 甘肃“问题隧道”网络传播分析

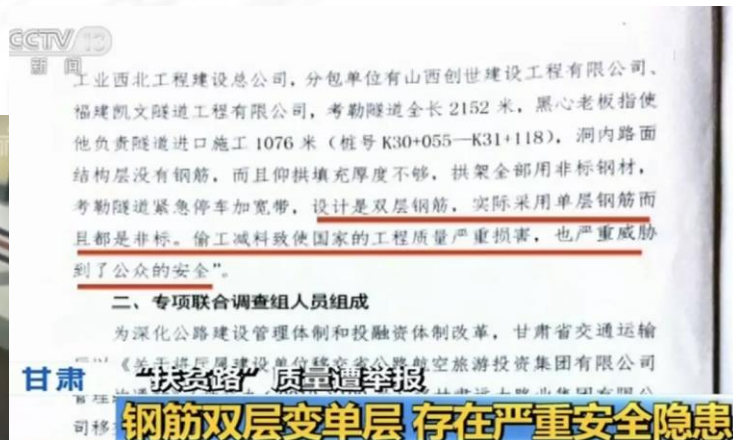
- 围绕关键词“问题隧道|16亿扶贫路|考勒隧道|折达公路|((甘肃+(质量问题|整改|曝光|刷涂料|偷工减料)+隧道))”，对2018/04/01 00:00~2018/04/09 23:59期间，互联网上采集到的179457条信息进行了深入分析。全网声量最高峰出现在2018/04/02 00:00:00，共产生108555篇相关讯息；
- 事件源头于2018/04/01 00:48分发布在微信上，题名为『怒！钢筋双层变单层，“整改”只是刷涂料，...』。后续报道主要来源于新浪微博、微信、人民政协报、机电之家、搜狐网等几大站点。总体来说，整个事件的发展趋势较为突出。

怒！钢筋双层变单层，“整改”只是刷涂料，这就是16亿建成的“扶贫路”？！

人民网 4月1日



总投资近16亿元的折达公路，是甘肃省专门修建的过的甘肃临夏回族自治州东乡县，属于国家级的贫穷，群众出门只能爬山或坐渡船。而这条公路的通的出行问题，对于经济发展也有非常大的作用。





# 事件走势

04月01日

[4月1日 0点]怒！钢筋双层变单层，“整改”只是刷涂料，这就是16亿建成的“扶贫路”？！[微信] 影响力：1

[4月1日 12点]【#16亿扶贫路偷工减料# 整改就是刷遍涂料？[怒]】国家投资近16亿的甘肃扶贫路，竟遭偷工减料！双层钢筋变单层，路基裂缝随处可见！隧道上方是大山，隐患不堪设想！整改就是刷了遍涂料...甘肃公路管理局副处长杨爱明称，我没空上去看去...甘肃交通运输厅领导不见记者，想待着就待着，不想待着就走...<http://t.cn/RndOusR>[新浪微博] 影响力：127375

[4月1日 22点]关于中央电视台曝光折达公路考勒隧道问题处理进展情况的通报[甘肃省交通运输厅] 影响力：4024

[4月1日 23点]甘肃回应扶贫路刷层涂料就算整改:立即启动问责[黄河新闻网] 影响力：2939

[4月1日 23点][24小时]甘肃交通运输厅发布通报 对考勒隧道质量问题进行处理[新闻联播] 影响力：157

[4月2日 8点]甘肃成立调查组调查折达公路考勒隧道问题[华龙网] 影响力：387

[4月2日 23点]【解局】甘肃官场到底出了什么问题？[搜狐网] 影响力：276

[4月3日 15点]豆腐渣作风比豆腐渣工程更可怕[中国农业新闻网] 影响力：224

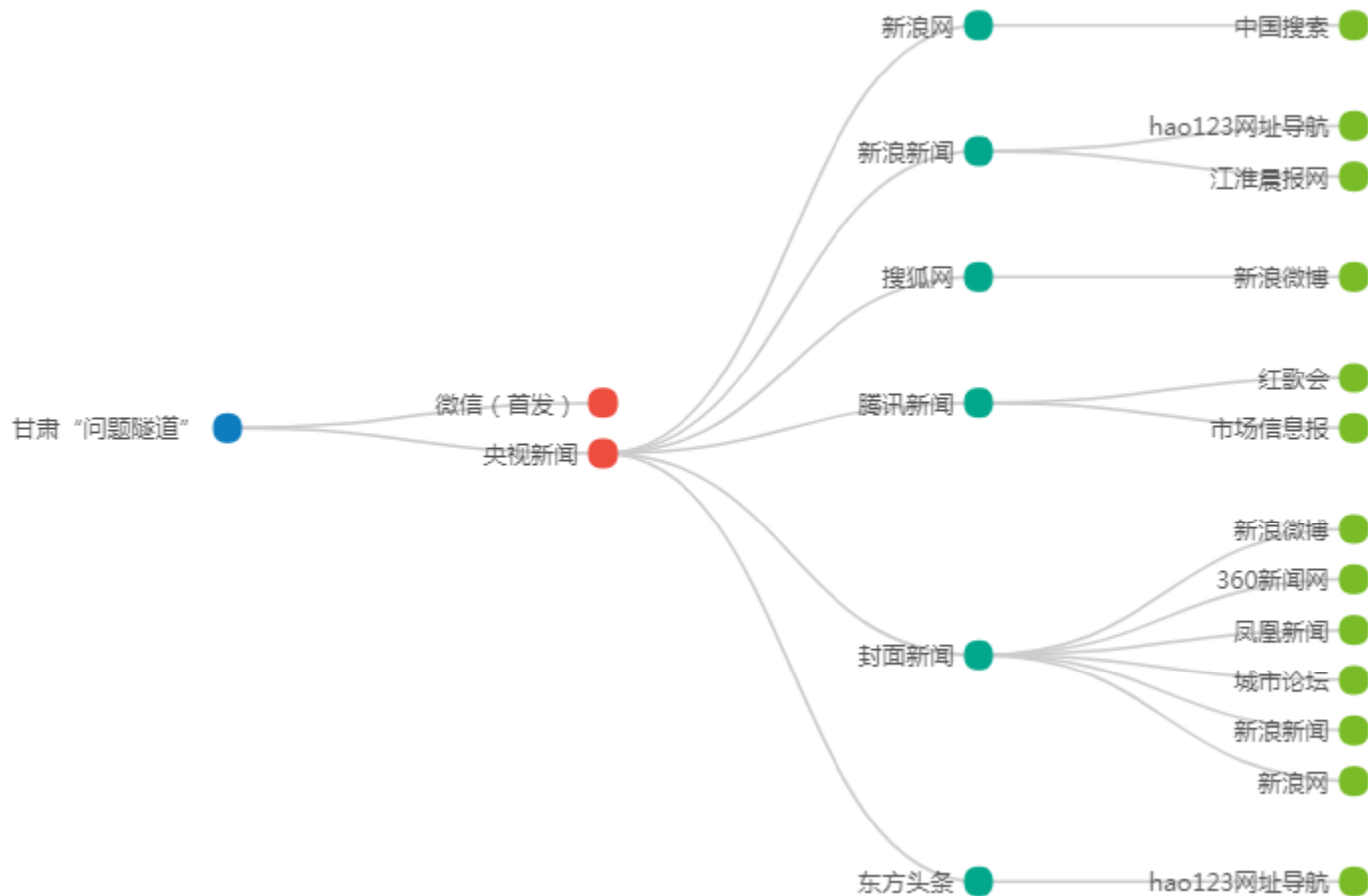
[4月3日 21点]国家道路及桥梁质检中心对甘肃折达公路全线进行全面检测[澎湃新闻] 影响力：849

[4月4日 10点]甘肃：零容忍查处问题公路腐败 排查全省交通项目[看山东] 影响力：309

完



# 传播途径



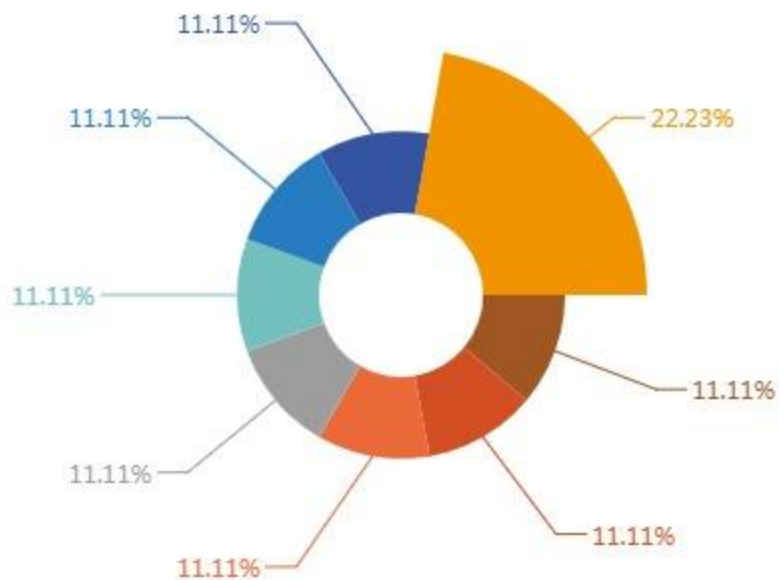


# 关键词云





# 微博观点分析



- 我们村的公路是年年修年年坏啊(22%)
- 冬天是我们这冷冻破管子也是情有可原(11%)
- 不偷工减料 当官的想收红包(11%)
- 这种问题 全国哪里不都是这样(11%)
- 国家应该对此类工程进行严抓(11%)
- 这次不好说 赶到风头上了 要被当鸡杀(11%)
- 这个新闻就是因为当地的村名举报了(11%)
- 公路局那个处长说的没错啊(11%)





# 輿情总结

- 综上所述，在『甘肃“问题隧道”』事件/话题中，
- 媒体主流报道为『国家投资近16亿的甘肃扶贫路』
- 网民主流意见为『我们村的公路是年年修年年坏啊』
- 应深入挖掘网民意见和情感倾向，识别事件传播过程中的意见领袖和主要信息来源，预测或追踪舆论走向，以便对不良舆论进行疏导

@央视新闻

【#16亿扶贫路偷工减料# 整改就是刷涂料？】国家投资近16亿的甘肃扶贫路，竟遭偷工减料！双层钢筋变单层，路基裂缝随处可见！隧道上方是大山，隐患不堪设想！整改就是刷了涂料...甘肃公路管理局副处长杨毅明称，“我没空去看去...”甘肃交通运输厅领导不见记者，“想待着就待着，不想待着就走...”... 展开全文



4月1日 12:35 来自 微博 weibo.com

95899 73340 112136

## 甘肃媒体:一条隧道非全部 没理由质疑整个甘肃官场

2018-04-03 20:24:58 来源: 兰州晨报(兰州)

▲ 举报

63323

(原标题:金城观 | 甘肃并不是你们说的那样)

兰州晨报

甘肃媒体: 一条隧道几个官员暴露问题, 不能代表整个甘肃官场

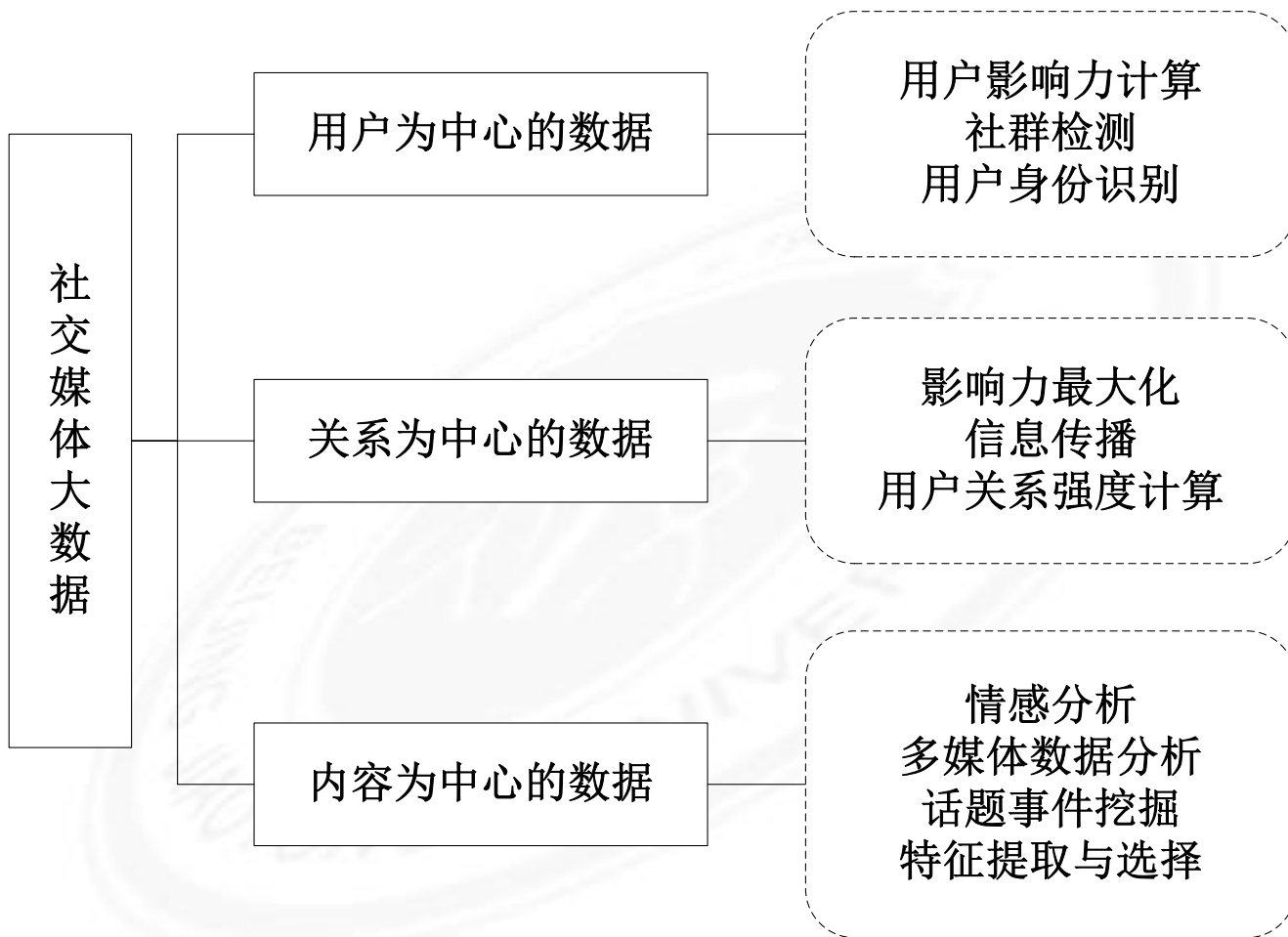
的质量问题曝光在众目睽睽之错的采访, 交通主管部门的几个

尔东/兰州晨报APP“掌上兰州”  
2018-04-04 10:53

A+



# 基于用户的大数据分析

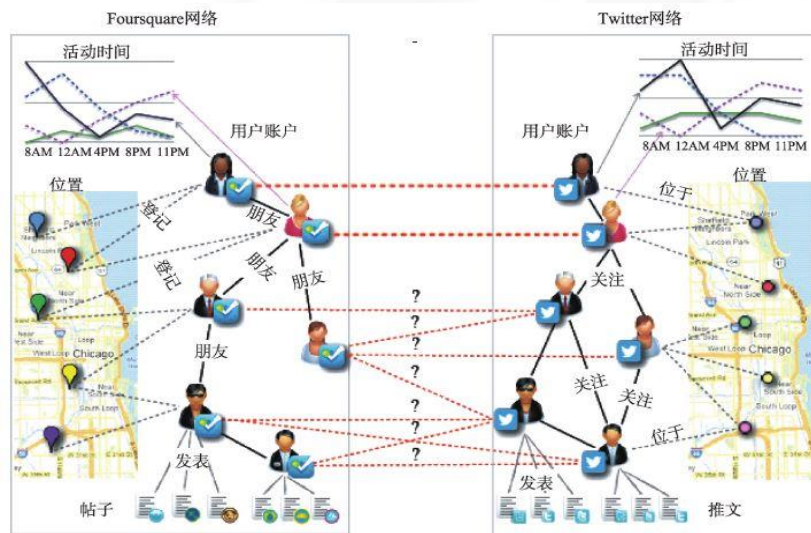




# 基于用户的大数据分析

## • 用户识别

—在线社交网络可看做异构信息网络，其中的信息通常包括时间、地点、人物、事件等，而用户往往同时存在于多个不同的社交网络中。由于异构的特点，导致同一个人在不同的网络中会呈现一定的差异，如何在此种情况下识别这个人的身份成为近年来异构社交网络研究的一个热点。





# 基于用户的大数据分析

## • 社群发现

–社群是指用户在某段时间内互动形成的具有稳定群体结构、一致行为特征和统一意识形态的个体和社会关系的集合。社群内部用户关系强度强，聚合强度大，而社群之间用户关系强度弱，离散程度大。

–社群挖掘的目的在于从用户的行为、群体结构和关系模式中发现潜在的规律。社群结构按照用户社会关系和对文本内容的兴趣度划分为两种：

- 1、以用户个体为中心的社群结构
- 2、以话题为中心的社群结构

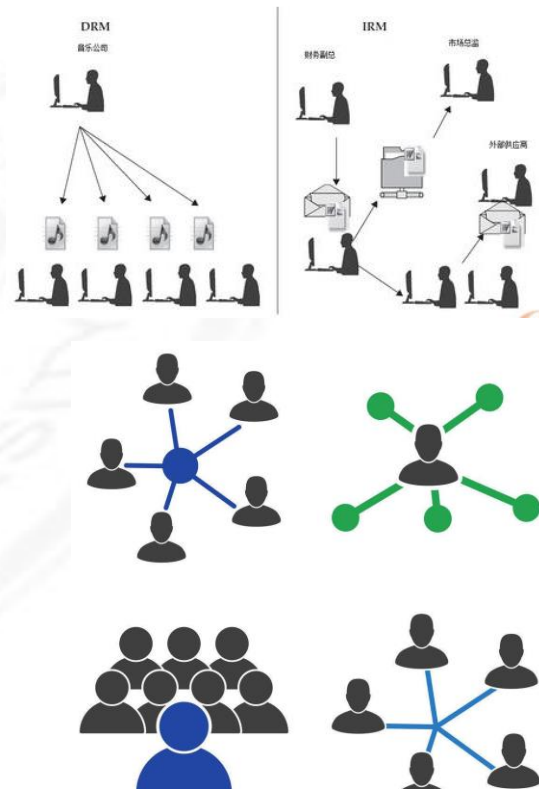




# 基于关系的大数据分析

表2 用户关系强度计算

度量方法	度量指标	网络结构
相似度计算	两节点的邻居重叠度	依赖
边介数	经过当前边的最短路径的总和	依赖
影响力图	弧的重数	依赖
隐含变量模型	描述内容的相似度与用户间的交互关系	依赖
时间模型	指数衰减模型	依赖







# 基于关系的大数据分析

## •信息传播

—用户关系强度的计算源于实际数据的传播模型，它们采用信息本身特性、用户关系、微博网络外部因素等多方面对信息传播进程建模，预测信息传播动态以及用户个体的传播行为。从整体出发，预测信息的扩散速度、范围、广度和深度等；或是从个体出发，预测用户个体传播某条信息的概率，进而研究整个社会网络的信息传播情况。

## •影响力最大化

—影响力计算是针对单个用户节点而言的，而影响力最大化问题涉及网络中的多个用户，考量集体的联合影响力，它利用信息传播模型聚集用户，使用户集合可以最大程度地影响其他用户，从而使信息最大程度地扩散。

—传统影响力最大化问题

—新型影响力最大化问题

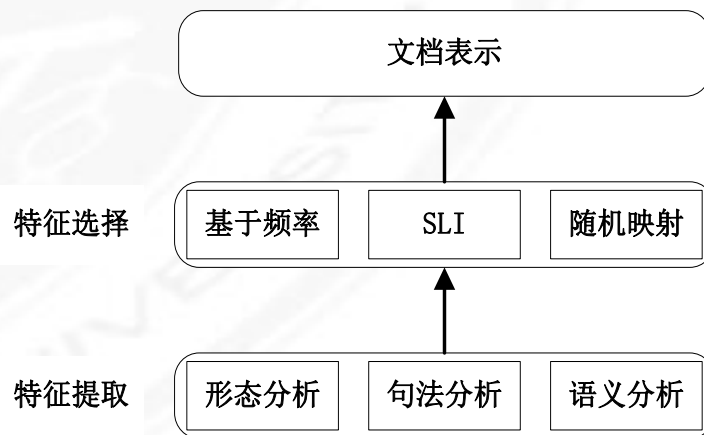




# 基于关系的大数据分析

## •特征提取与选择

– 收集到的原始文本组织松散，直接用于文本分析会影响分析的准确性。预处理就是采用特征抽取和特征选择的方法将文档组织成固定数目的预定义类别。





# 基于内容的大数据分析

## •话题事件挖掘

– **事件**是指在特定的时间和地点下发生的有前因和后果的事情，而**话题**是指由所有直接相关事件构成的大事件。话题挖掘的主要任务是话题检测与跟踪，采用历史事件追溯检测和在线新事件自动识别方法，对此已有大量研究，尤其针对完整新闻报导和博客的话题检测已取得了一些成绩。

– 然而，由于微博格式复杂，内容简短，用语不规范等特点，TDT技术不能简单应用到微博。

## •话题模型

## •话题摘要

## •话题的检测与跟踪

## •多媒体数据分析

## •地理位置信息分析

## •社交媒体的动态性和时效性分析

## •社交媒体大数据中存在的深层语义挖掘

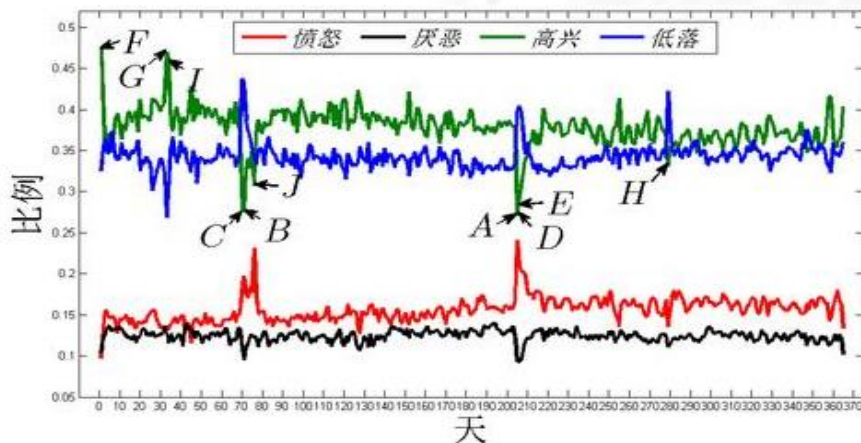


# 基于内容的大数据分析

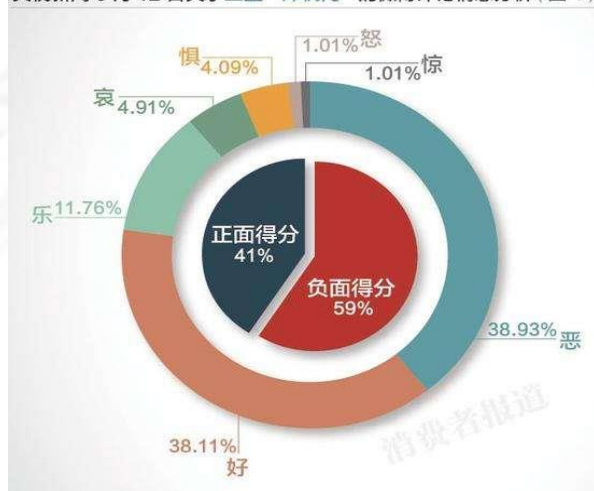
## •情感分析

—情感分析也叫意见挖掘，旨在依据意见目标从语料中识别和提取特定主题的属性、要素和隐含的主观信息。意见目标通常称作实体，可以是人物、事件或话题，与要素和子要素相关联，每个要素都有其自己的一套情感属性。

—微博情感分析可以提取不同领域的公众情绪和意见，可以确定民意调查的影响，有效解释和描述政治事件，预测股票趋势等。



央视新闻 9 月 12 日关于三星“炸机门”的微博评论情感分析 (图 1)





# 大数据与社交媒体的融合

- 社交媒体大数据的未来挑战





# 社交媒体大数据的未来挑战

- 挑战
  - 信息传播效应刻画
  - 影响力计算
  - 特征提取与选择
  - 微博新闻挖掘
  - 社会媒体大数据融合
  - 跨语言情感分析



# 社交媒体大数据的未来挑战

- 信息传播效应刻画

- 社交媒体网络中信息传播效应的刻画是一个复杂的问题，它受到信息自身因素、社会因素和网络外部因素的综合影响，并且用户本身的属性与信息本身的属性也相互影响，准确全面地反映信息传播效应已成为关键。这一问题的解决还依赖于影响力、用户关系强度和传播规律。



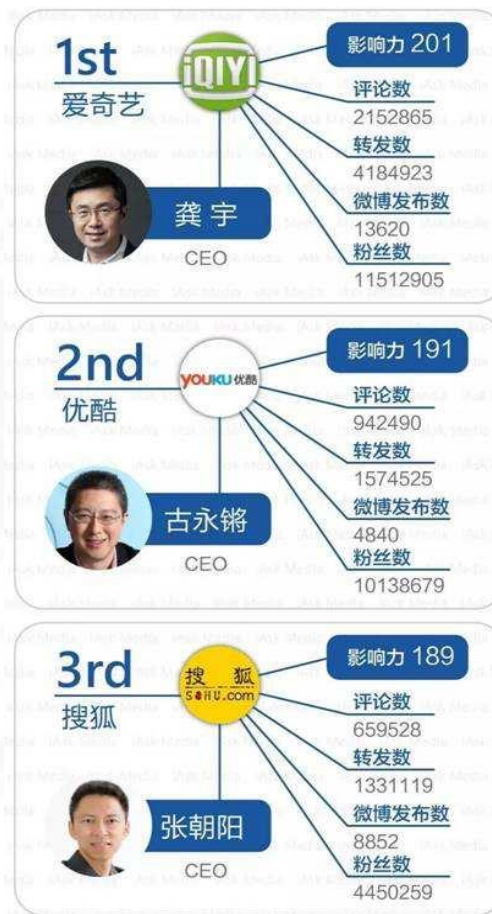




# 社交媒体大数据的未来挑战

## • 影响力计算

- 基于关系分析的一个具有重要商业价值的研究方向是影响力计算和信息传播的最大化问题。其中信息传播的最大化问题的全局最优化被证明是**NP难问题**，对于大规模的社会网络，目前只能采用一些优化算法获取近似的较优解，并且对于影响力最大化问题目前的最佳解决算法也只处理了百万级规模的社会网络。而目前微博网络节点过亿，如何在微博网络中快速计算出固定数量的最有影响力的节点集合还有待进一步探究。

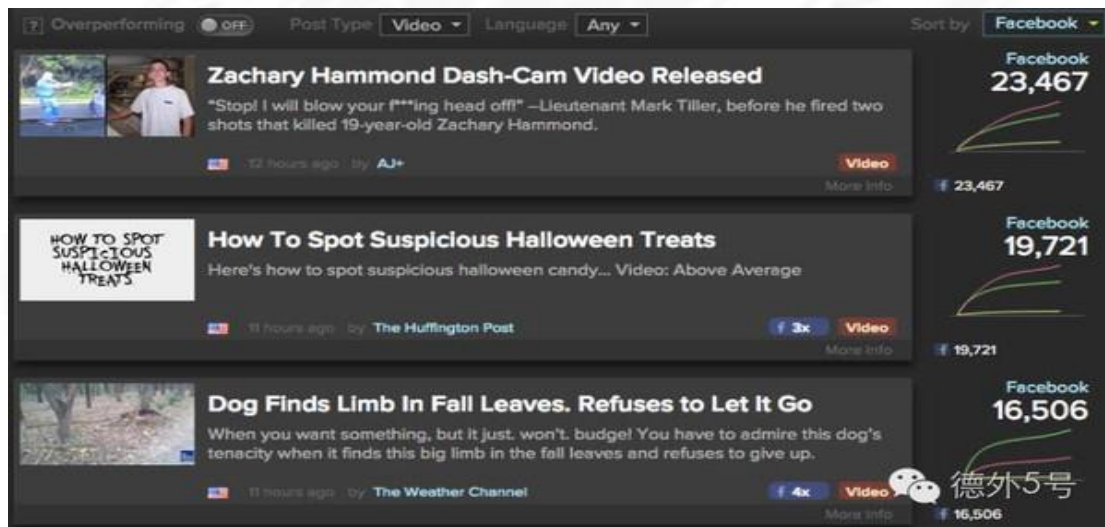




# 社交媒体大数据的未来挑战

- 特征提取与选择

- 针对传统数据的特征提取与选择方法已有很多，但是不利于处理低频词和发现新特征，而这种情况在微博数据中大量存在。与词频模型相比，序列模式挖掘保持了词的顺序并可以捕捉潜在的语义，更能解释话题。但是采用模式挖掘的两大挑战是：大量冗余模式的产生和长模式的低支持度问题。





# 社交媒体大数据的未来挑战

- 微博新闻挖掘
  - 如何在线实时处理这种社会化的短文本流?
  - 如何识别新闻话题?
  - 如何实时检测新闻事件?
  - 如何判断事件的连续性?
  - 如何挖掘这种动态的关联演化性?
  - 如何从海量博文中提取有意义且易理解的微博话题?
  - 挖掘到的新闻以什么形式呈现?
  - 如何设计针对微博的动态新闻集成系统?

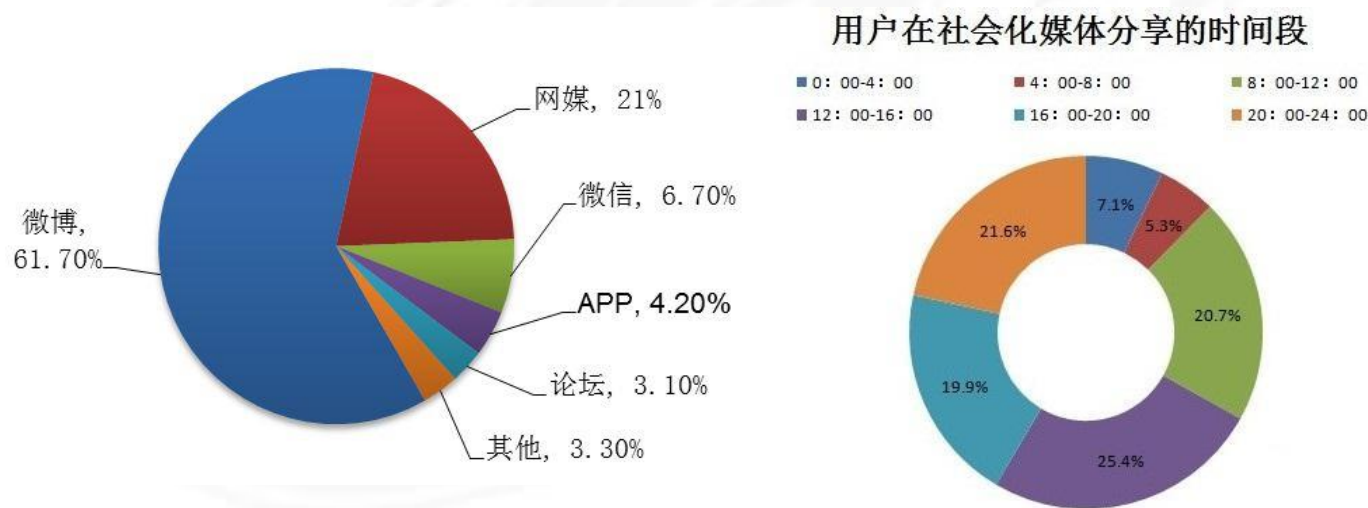




# 社交媒体大数据的未来挑战

## • 社会媒体大数据融合

- 随着社会网络服务的发展，用户在社交互动中加入了多种服务，并收集了大量的信息。因此，如何整合分布式社会网络，进而对各种社会媒体数据源进行融合，为知识的挖掘提供更好的数据资源已经成为亟待解决的问题。





# 社交媒体大数据的未来挑战

- 跨语言情感分析

- 挖掘情感是为了体现商业价值，目前大数据向跨语言融合迈进，相应的情感分析也向跨语言情感分析发展。但是，语言的不同体现在语言特征、要素分布的不同，语言间关联的障碍使得跨语言情感分析成为更大的挑战，这是目前亟待解决的问题。

