以主体为中心的微博计算方法[1]

——微博计算微革命：从“信息”中心到以“人”为本

张华平 商建云 赵燕平
北京理工大学
北京海淀区中关村南大街 5 号 100081

摘要：微博逐渐超越电视、新闻、论坛，成为新的具有革命性的社会化媒体，注册用户数超过 2 亿，使用率达超过 50%，对微博进行计算挖掘分析已经成为学界与产业界共同关注的课题。微博内容的碎片化与网络主体化特征日益凸显，以信息内容为中心的传统计算模式存在本质缺陷，其信息数量巨大，处理效率低下且效果很难满足实际需求。本文针对以微博为对象的分析挖掘，提出了“以人为本”的微博计算模型，即以微博主体为微博计算的主要对象，研究微博博主个性化表示模型，博主情绪感知算法、及微博内容分析等关键技术，在微博计算方面取得了较好的结果，本文的创新在于突破了纯粹内容分析的局限，更好地适应了微博计算的需求，是微博计算的方向所在。

关键词：微博计算；个性化建模；主体行为模式挖掘；数据可视化
中图分类号：TP139.1
文献标识码：B

Microblog Computing Focused on Social Entities

-- Micro-evolution in Microblog Computing: focus from information to entities

Abstract: The number of microblog user reaches over 200 million and usage rate is 50% . Microblog is becoming revolutionary social media instead of TV, news and forum. Caculating and doing data mining from microblog is becoming common hot

[1] 本文工作得到了国家自然科学基金面上项目（项目号：61272362）、新疆自治区高新技术计划项目（项目号：201212124）支持。张华平(1978 年--)，男，江西鄱阳. 副研究员，博士，研究方向为：微博计算、自然语言处理、信息检索。

topics for the academia and industry circles. Because of content segmentation and entity personalization in microblog, the traditional computing schema focused on content computation is inherent defective. The amount of mciroblogs is huge, the content based computation is ineffective and hard to meet the requirement. This paper brings forth a novel computing schema based on social entities in microblog computing. And then research on microblog entities personality modeling, emotion detection, and further develop content analysis based on social entities. Such computing schema focused on social entities achieved better performance than information centered computing with microblog analysis and mining. The contribution is break out the pure content analysis, and social entities centered computation is more suitable for microblog computing.

Key Words: Microblog computing, personalized modeling, data visualizing, behavior model mining

1. 引言

微博客（microblogging 或 microblog，简称微博）起源于美国的 Twitter (twitter.com)，是一种允许用户及时更新简短文本（通常少于 140 字）并可以公开发布的博客形式。目前，中国微博注册用户数达到 2.5 亿，使用率达到 48.7%，仅新浪用户每日发博量超过 1 亿条。其历史虽然只有二年的时间，却用一年时间发展成为近一半中国网民使用的重要互联网应用[1]。微博信息传播速度快，影响范围广，实名用户数量庞大，与真实社会相互交融，密不可分。每天有数以千万计的微博用户通过微博表达各自对社会生活各个层面的观点，表达个性化的感情，所涉及的内容保罗万象。微博已经逐渐超越电视、新闻、论坛，成为舆情话题产生和传播的主要场所，形成社会舆情的引爆点与主战场。可以说，网络信息和社会信息的交融对社会的直接影响越来越大，甚至关系到国家信息的安全和社会的长治久安。

在 2009 年，Twitter 引爆摩尔多瓦颜色革命并进行活动串联，导致重大动乱，成为微博第一例参与社会重大活动的案例；2011 年风起云涌的阿拉伯之春、伊朗骚乱，以及震惊世界的伦敦骚乱等，也处处渗透着微博的力量。传统媒体占据舆论制高点的地位正在逐步削弱，而要在重大公共事件中起到引导与辟谣的实效，已经到了不能回避微博、社交网络这类既经济又及时的传播工具的时代了。2011 年 2 月，美国国务卿希拉里在乔治华盛顿大学就互联网自由问题发表讲话。她强调以 Twitter 为代表的社交网络，是美国政府的一种重要战略力量。美国国防部长罗伯特•盖茨 2009 年 6 月表示，Twitter 等在伊朗德黑兰抗议活动中起到重要作用的社交网络是“美国重要的战略资产”。

本文第二节将阐述传统以“信息”为中心的微博分析计算所面临的特有挑战，随后重点介绍以主体为中心的微博计算方法，给出微博计算的研究框架，最后介绍了我们在微博计算的研究进展。

2. 以“信息”为中心的微博分析面临挑战

对于传统常规文本的分析，用以信息为中心的分析方法能够解决信息挖掘的目的，但对微博类短文本的分析任务更为艰巨。到目前为止，能收集到为数不多的有关微博信息分析的论文多是研究英文 Twitter 的，研究的内容也多是情感分析的。文献[2]首次提出将 Twitter 作为情感分析与观点挖掘的语料库，认为 Twitter 相对于传统网络应用形式，更适合于情感分析，并通过在收集到的 Twitter 语料库基础上进行了语言分析。文献[3]在针对 Twitter 的消息文本进行情感分析后，能较为准确地预测出投票结果，并提出可作为社会投票调查的一种有效的替代方法。有研究者针对 2008 年到 2009 年间的消费者信心与政治观点，发现与同时期的 Twitter 消息中出现的情感词词频正相关（在不同数据集的实验表明相关度高达 80%），因此能够通过大规模的消息情感分析，把握各类话题的总体发展趋势。在与传统常规文本情感分析的对比上，文献[4]在对英文 Twitter 的研究中发现，短文本在情感分析方面反而更加精准，其发现有一定的语言学依据，但其结论是否适用于中文微博，还有待进一步验证。

在对国内外学者的研究成果分析的基础上，本文认为新型社会网络的信息计算遇到了前所未有的挑战，主要包括：

1）微博社会网络包含的信息内容短小但规模巨大，如微博每条最多 140 字，每天原创的微博数千万条，导致单条内容的分析极其困难，而总体计算的代价极大；

2）微博社会网络内容不规范，语言口语化严重，且有上下文背景，单条内容很难被完整正确的分析；

3）传统静态网页可追溯可脱机计算，而微博社会网络的信息快捷，稍纵即逝；对社会网络信息的计算需要足够高效；

以“NLP”（自然语言处理）为例，可以通过微博搜索到如下结果，不同微博对“NLP”存在多种歧义理解，从内容上很难区分和理解。我们在人工综合微博博主的各种背景资料之后，会发现“NLP 学院”是专门从事心理学培训相关的，因此，一条微博中的 NLP 居然表示的是“neuro linguistic programming”（即身心语言程序学，NLP 是关于人类行为和沟通程序的一套详细可行的模式，是很热门的心理学和成功学的培训课程）；同样，只有在了解博主刘知远 THU 是清华大学计算机博士之后，我们才能真正理解另一条微博中的 NLP 含义为“natural

language processing”即自然语言处理。

从上面的分析与例证中，我们不难得出一个结论：如果将社会网络计算的着力点放在单条内容的深度理解上，其性能和效果都存在本质上的缺陷；因为常规长文本分析将不适于微博类的情感分析，已有的以信息为中心的方法存在以下 3 个方面的不足。

（1）没有考虑作者或发言人等信息主体的个性化背景知识。

（2）由于情感分析客观上和观点持有者有很大的关联性，仅靠分析文本内容本身是远远不够的。

（3）主体属性的缺失导致情感分析的客观性和科学性大打折扣。

而如果以微博网络主体为中心，综合考虑网络主体的多维信息（包括主体的基本资料、发布的内容信息、个人关系网及用户的社会行为）并在此基础上综合利用自然语言处理、社会网络分析、数据挖掘、信息传播等诸多交叉学科知识，对微博社会网络数据进行挖掘与分析，才能引领正确的社会网络计算发展方向，将对微博社会网络的分析挖掘提供理论基础，为社交网络分析以及新型网络舆情监测等诸多应用奠定坚实基础。因此，本项研究将以微博网络主体的个性化表示为切入点。网络主体的个性化表示是社会网络的一个基本科学问题，以网络主体为中心的新型计算模式，将对传统的信息内容为中心的模式进行革新，且必将推动网络科学的发展。

3. 以“人”为本的微博计算

虽然微博是最近几年才兴起的，且这方面的研究起步较晚，但其具有更广阔的应用前景。2011 年社会心态蓝皮书表明网民上网的第一站登录微博的比例将近 20%，直逼即时通信工具和电子邮件；微博对门户网站流量的贡献度在逐步增加，微博导入用户是网站的优质用户。由此可见，微博带来的微革命已经悄然而至。由于越来越多的用户乐于在互联网上分享自己的观点或体验，这类评论信息迅速膨胀，有可能会成为舆论快速传播，形成重大事件。

然而，如何从这些微博网络信息中挖掘出重要的信息战略资源不是简单能够做到的，由于其分析的内容短小且不规范，一般很难充分利用上下文信息，必须综合多个学科，对微博进行建模、分析，才能综合挖掘出其后面所隐含的信息或知识，从而成为重要的“战略资产”。为此我们对微博这种内容的分析综合问题率先提出了以“人”为本微博计算的概念。

所谓微博计算是指以微博类个性主体为中心，综合计算语用学、社会网络分析、传播理论等学科，对微博进行建模、分析与应用等计算的研究方向。由于微

博数量的指数级的增长，对社会的影响越来越大，在做微博计算时会遇到很多问题，需要找到解决方案并且不断进行优化。

为了解决上述的问题，我们提出了从以信息为中心转为以人为中心的研究内容框架。也是微博计算所包含的主要内容。

首先通过各种方法获取社会网络主体数据进行分类，从而划分出基本属性数据、内容与评论数据以及关系数据；以此为基础，在图 1 中的 4 个大的支柱计算后综合出个性化的表示模型进行后续的诸如好友推荐等服务；这 4 个大的支柱加上相似度计算构成了微博计算的核心内容。

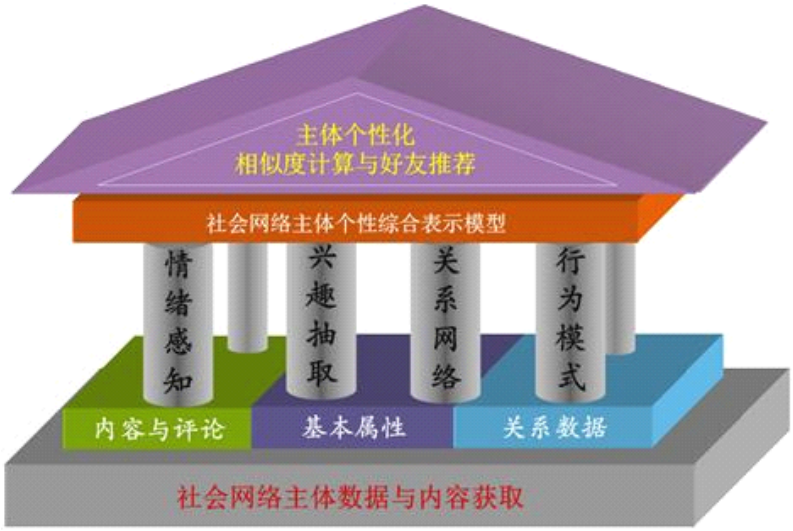

图 1 研究内容框架示意图

如上图所示，微博计算的核心内容概括为以下 5 个方面。

(1)网络主体的个性化综合表示

综合微博网络主体的基本属性、行为模式、个人兴趣、情绪特征、以及关系特征等维度的计算结果，借鉴面向对象的思路，研究社会网络主体的表示模型。

网络主体的综合表示模型如下：

$$
\begin{aligned}
E = E[\ & info, \\
& Pattern(content, info), \\
& KeyExtract(content), \\
& EmotExtra(content), \\
& RelationMining(relation, content) \\
]\ &
\end{aligned}
\qquad \cdots\cdots (1)
$$

（2）社会网络主体的行为模式挖掘

根据社会网络主体的原创发帖、回复与评论、转发以及发送图片等历史活动过程，采用数据挖掘等相关的技术手段，挖掘出（1）式中对应的社会网络主体

的行为模式向量，从而分析出其个性化的行为特点。

（3） 主体个性相似度计算与推荐算法

在网络主体个性化向量表示模型基础上，可计算不同网络主体之间的相似度，而在内容向量上，则采用信息检索传统的向量相似度计算策略，计算公式如下：

$$\begin{bmatrix} 1 & Corr(X_1, X_2) & \cdots & Corr(X_1, X_n) \\ Corr(X_2, X_1) & 1 & \cdots & Corr(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ Corr(X_n, X_1) & Corr(X_n, X_2) & \cdots & 1 \end{bmatrix} \quad \cdots\cdots\cdots (2)$$

对基本信息可采取相似度规则匹配的方法，可以针对微博的具体应用采用不同个体之间的相似度，并研究好友推荐的相似度排序算法。实现高效的好友推荐。可以采用决策树的方法综合考虑不同的要素，在具体要素方面进行相似度比较，可以得到相似个体的聚类。

（4）基于内容的个性化兴趣与情绪感知

情感分析（Sentiment Analysis），又称情感倾向性分析、意见挖掘（Opinion Mining）或情感分类（Sentiment Classification），最早可以追溯到上世纪九十年代，它试图用计算机实现从文本的内容中提炼出作者的情感方向的目标。通过情感分析，可以明确网络传播者所蕴涵的感情、态度、观点、立场、意图等主观反映情感（如乐、好、怒、哀、惧、恶、惊）。可以找到个体的兴趣点。

根据社会网络主体发表的所有信息内容，采用自然语言处理技术，抽取文本内容的个性特征词，并分析每条消息所传达的情绪色彩，研究主体个性化的兴趣提取算法，并从每条消息的情绪中，综合研究得出网络主体的情绪特点与情绪波动规律，实现情绪感知。采用标签云的方式表示与呈现。

（5）个人关系网的特征分析

主体个性化的微博情感分析中个性化因素可以形式化定义为：

$$U = U(info_u, reliability_u, influence_u, relation_u, history_u) \quad \cdots\cdots\cdots(3)$$

基于主体个性化微博内容情感分析计算公式如下：

$$S = SentimentAnalysis\ (C, U)$$

$$= \sum_{sent_i \in C} [\lambda_i sent_i \times (\alpha Info_U + \beta Relation_U) \times Influence_U \times e^{Reliability_U}]$$

其中：$sent_i$为情感词，$Info_U$为用户U的基本信息

$Relation_U$为用户U的关系网，$Influence_U$为用户U的影响度

$Reliability_U$为用户U的可信度，λ_i，α，β为调节参数。 $\cdots\cdots\cdots (4)$

对于各主体个性化要素与内容分析的结合问题，拟采用随机森林集成学习法和其他的机器学习相结合的方法，通过实验确定各因素的调节参数，并得到个性化的内容情感分析指标值。

根据社会网络主体关注列表与被关注列表，以及与好友互动的记录，并利用前述的个性化要素和个性化内容情感分析指标，采用社会网络分析方法，引入个人关系网的特征提取算法，分析关系网中的关键人物、亲密好友，以及主体所属社区的自动分类。用可视化图形展示出个体与个体的关联关系等。成果展示中我们给出了一个实例。

4.微博计算已有研究进展

最底层为社会网络主体数据与内容的获取。早在2009年我们就开始了中文微博的研究，北京市网络控制办公室的微博监测分析已经开始与课题组合作。采用基于浏览器模拟仿真方法实现了微博信息采集，目前已经获取了微博主体语料库6000万条，都具有一定的“粉丝”数，剔除了大量机器自动生成的用户信息，部分博主样本数据见图4。已经覆盖中文微博社区的主流活跃用户（活跃用户约占总注册用户的5%），各类信息齐全。为了提高效率，采集了以下三种关键策略。
1）定向垂直采集：定向垂直采集特定敏感的微博人群，确保信息的及时性；
2）元搜索主题采集：针对特定内容，充分利用微博自带的搜索功能，采集主体相关的的信息内容和信息要素，如内部id、性别、家庭住址、粉丝数目、个人摘要、微博数量、关注数量、博客地址、教育情况、工作情况、是否认证、生日。以最小的成本确保信息的覆盖面；
3）并行采集：针对同构的海量采集任务，采用多机器并行处理，基于云计算架构，实现多任务多策略的并行采集。通过公开采集与抽取从新浪微博、腾讯微博中获得。为了推进微博计算的研究，我们通过在自然语言处理与信息检索共享平台上(www.nlpir.org)，予以公开公布其中的经剔除了大量冗余与机器粉丝后的20万条数据。

2012年2月14日，课题组免费发布了NLPIR微博语料库，其中包括100万微博博主语料库、100万微博关注关系语料库、100万微博内容语料库，是目前已知的第一家公开共享的微博语料库，广受学术界与产业界的认可。课题组组长张华平博士作为发起人，创建了第一个专门研究微博的社区“微博计算”，目前已经吸引了国内外众多的研究者与爱好者加入。另外课题组在新浪微博发布了微博个性热词云服务，主要用于分析博主的个性化兴趣，并提供不同微博博主的兴趣相似度比较应用，目前已经有1万多用户在使用该服务。

在获得了这些数据的基础上在5个方面取得了研究成果。

1） 社会网络主体的行为模式挖掘

以微博博主的发博时间为例对其作息时间进行了行为模式的挖掘。

作息时间以行为分布矩阵公式1表示，其中每列表示的是每天24小时的时

段，每行表示的是周一至周日，实际的数据表示的是在该日该时段社会网络主体的活动概率，如图 2 所示。

时段\周几	周一	周二	周三	周四	周五	周六	周日	时段边缘分布	
0	0.03%	0.14%	0.26%	0.11%	0.11%	0.34%	0.31%	1.31%	
1	0.20%	0.26%	0.00%	0.03%	0.00%	0.20%	0.11%	0.80%	
2	0.00%	0.09%	0.03%	0.00%	0.00%	0.00%	0.28%	0.40%	
3	0.03%	0.14%	0.00%	0.00%	0.09%	0.06%	0.03%	0.34%	
4	0.09%	0.11%	0.06%	0.00%	0.06%	0.00%	0.03%	0.34%	
5	0.06%	0.00%	0.00%	0.00%	0.06%	0.09%	0.14%	0.34%	
6	0.09%	0.14%	0.34%	0.23%	0.20%	0.20%	0.17%	1.36%	
7	0.45%	0.54%	0.60%	0.60%	0.37%	0.48%	0.54%	3.58%	
8	0.80%	1.14%	0.97%	0.82%	0.71%	0.68%	0.28%	5.39%	
9	0.94%	0.74%	0.68%	0.85%	0.85%	0.68%	0.54%	5.28%	
10	0.26%	1.28%	0.74%	0.62%	0.91%	0.65%	1.11%	5.57%	
11	0.40%	1.19%	1.28%	1.16%	1.36%	1.08%	0.74%	7.21%	
12	0.45%	0.45%	0.97%	0.74%	0.85%	0.74%	0.65%	4.86%	
13	0.45%	0.60%	0.71%	0.60%	0.91%	0.62%	0.74%	4.63%	
14	0.71%	1.42%	1.68%	1.45%	0.99%	0.71%	0.60%	7.55%	
15	0.88%	1.08%	1.82%	0.37%	1.05%	0.68%	0.43%	6.30%	
16	1.11%	0.82%	0.85%	0.97%	1.33%	0.80%	0.74%	6.62%	
17	0.54%	1.39%	1.33%	0.85%	1.33%	0.94%	0.54%	6.93%	
18	0.85%	1.19%	1.16%	1.16%	1.02%	1.25%	0.68%	7.33%	
19	0.57%	0.57%	1.08%	0.99%	0.88%	0.71%	0.48%	5.28%	
20	0.51%	0.57%	0.77%	0.45%	0.54%	0.48%	0.71%	4.03%	
21	0.51%	0.60%	0.80%	0.99%	0.71%	0.60%	0.37%	4.57%	
22	0.88%	0.74%	1.16%	0.71%	0.62%	0.99%	0.77%	5.88%	
23	0.48%	0.48%	0.51%	0.62%	0.71%	0.85%	0.45%	4.12%	
周几边缘分布	11.27%	15.67%	17.77%	14.34%	15.67%	13.83%	11.44%	100.00%	2.790334059
								4.262131289	6.983034994

图 2 微博博主的作息行为分布矩阵图

基于这种行为的模型表示，利用协方差矩阵及信息熵等手段，采用可视化界面，我们对人物’潘石屹”2 年的微博活动规律进行了跟踪分析，可以计算出人物“潘石屹”的行为模式的可视化表示。

采用同样的模型，我们拟对行为模式挖掘进行如下更进一步的研究工作包括：

（1）不同维度的对比：实现每日不同时段，每周不同日，以及每年不同月份等一系列的行为模式挖掘；

（2）社会属性判别：与社会学相关的知识结合，从行为模式中计算对象“工作日与周末”的对比，判断其“是否有闲”，进一步研究推导出网络主体的职位、工作性质、经济状况乃至政治倾向性；

异常行为发现：利用与已有行为模式的对比，可以侦获网络主体的异常行为，如异于往常的频繁活动，往往是针对突发事件进行危机公关，分析此时段的内容，可以高效准确地对突发时间实现提前预警；而偏离正常的静默期，很大概率上是网络主体的非常态变化。

根据社会网络主体的原创发帖、回复与评论、转发以及发送图片等历史活动过程，采用数据挖掘等相关的技术手段，研究社会网络主体的行为模式，从而分析出其个性化的行为特点。

进一步的还可以做以下的分析。

（1） 工作日与周末行为模式的比对，推测其经济状况，并进一步分析其

身份。

（2） 引入原创率、图片率、客户端、情感识别等维度

（3） 引入 big-seven 心理学模型，分析微博博主的情绪心理、性格，推测其民主政治态度

（4） 僵尸与水军自动识别：内容原创度，行为模式特征。

依据个体的行为模式，实现了群体搜索与监控。

2） 基于内容的个性化兴趣与情绪感知

在社会网络环境下，主体的言论与转发的内容往往透露了作者个性化的兴趣与情绪变化。博主的一举一动一言一行，看似偶然，偶然背后有必然的个性特征。

我们根据分词标注结果、上下位的丰富程度、词的文档分布及时间分布规律等因子，从消息内容中提取用户常用的热门关键词，识别新出现的词汇及短语，并与已有其他用户的内容进行对比，综合提炼出用户的个性化兴趣，并采用标签云的方式表示与呈现，标签云示例如下。

在情绪感知方面，主要是引入已有的情感识别的相关技术，融合行为主体的背景知识，宏观上得出主体的情绪变化趋势，进行综合的情绪研判。

图 3“张华平博士”的 2011 年 9 月 2 日个人兴趣标签云图

图 3 为张博士在 2011 年 11 月的研究兴趣标签云图，图的右侧分析的是张博士的博客中所用到的词的数量分布；依据这个数量分布以图的方式表示在图的左侧，表明到此时此刻他的研究兴趣标签，标签大的表明其对此方面的兴趣多，从图 3 可以看出他的最主要兴趣是网络，同样的方法抓取他 2012 年 2 月这个时间点上研究兴趣的云图，表明他对微博的关注已经超出了对网络的关注。依据这样的方法可以对主体进行特征演化分析。

除了对单一个体进行分析外还可以进行相关分析，找出二个不同的主体之间的共同兴趣。图中展示的是张华平博士与罗家德的共同兴趣是“社会”。

3） 主体个性相似度计算与推荐算法

图3是对一个主体的云图表示，图4则是对二个主体最近的200条微博进行了分析的云图表示。从图4中可以发现所分析的二个主体张博士和罗家德都对“社会”的关注比较多，根据微博中的总体综合表示模型，可以分别得到两个主体发表的

每条微博的特征向量形成特征矩阵所计算出的相似度微博相似度为19.35%，也就是说这二个主体的博客中对社会有着类似的兴趣。通过这样的计算可以分析出有相似兴趣的个体群，从而可以进行主题的提介服务。

图 4 张华平博士与罗家德的个人兴趣相似度计算标签云图

4） 个人关系网的特征分析

将特征提取算法引入到个人关系网分析中，分析出关系网中的关键人物、亲密好友，图 5 中的人物就是分析出的一个个人关系网可视化图形展示，说明这些主体有着某种联系，如共同的爱好等。有了这样的分析数据，再依据分析结果进行所属社区的自动分类。

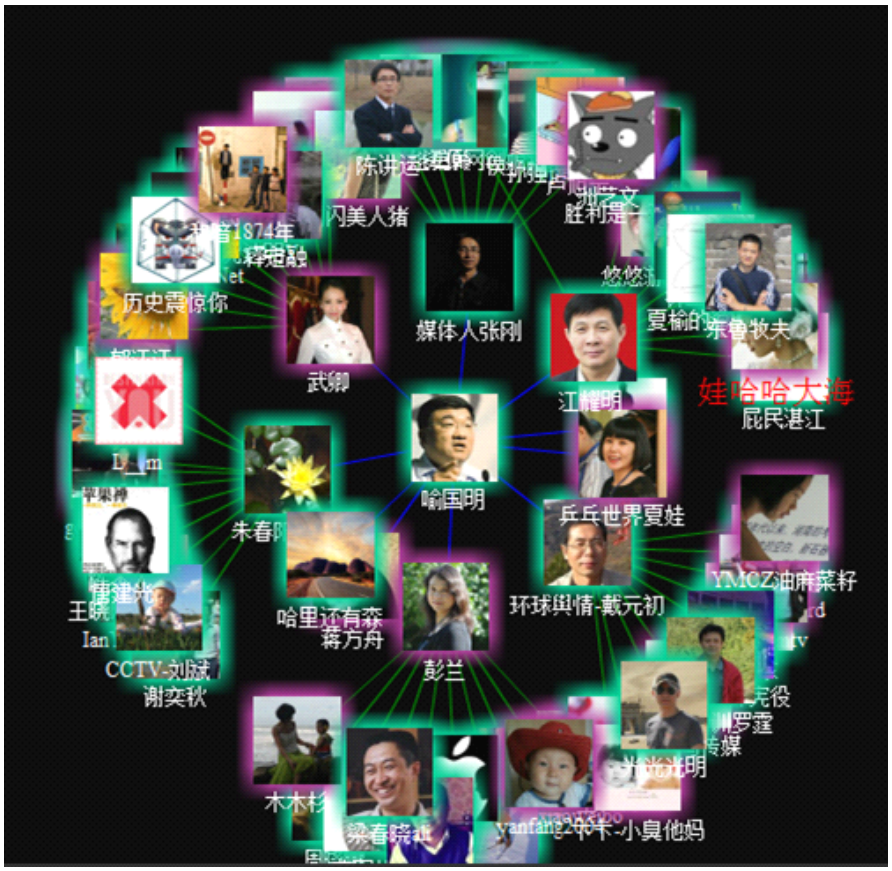

图 5 人民大学喻国明教授的关系网展示

5） 追踪监测研究成果

此外还对监测方面做了大量工作。实现了 ELINT 网络舆情挖掘系统，具有对微博话题人物监测、**谣言及辟谣监测以及对敏感问题追踪**的功能。

我们还研究了微博的革命性特征，并以舆情为例加以说明。在 2009 年，Twitter 引爆摩尔多瓦颜色革命并进行活动串联，导致重大动乱，成为微博第一例参与社会重大活动的案例。分析其演变过程可知有三个阶段，分别为线索渗透期、网络扩展期以及社会爆炸期。由此可知如何对其进行分析从而对舆情进行预警，对社会稳定是有重大意义的。

4. 结论

本文对对微博情感分析与感知的特殊性进行了论述，总结了了研究的难点与挑战。对这些难点和挑战问题给出了解决方法和模型框架。并将一些研究成果进行了介绍。

参考文献

[1] 第 29 次中国互联网络发展状况调查统计报告, 中国互联网络信息中心[DB/OL], http://www.cnnic.net/dtygg/dtgg/201201/t20120116_23667.html, 2012.

[2] Pak A.[J].Paroubek P. Twitter as a Corpus for Sentiment Analysis and Opinion Mining[C]//Proc. of Language Resources and Evaluation Conference. Lisbon, Portugal: 2010, pp.1320–1326.

[3] Brendan O[J]., Ramnath B., Bryan R. and Noah A. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series [C]// Proceedings of the 4th International AAAI Conference on Weblogs and Social Media, 2010.), 122-129

[4] Adam Bermingham[J] and Alan F. Smeaton: Classifying Sentiment in Microblogs: Is Brevity an Advantage? [C]// Proceedings of the CIKM, 2010: 1833-1836.