中国科学: 信息科学 2015 年 第 45 卷 第 6 期: 783–795
www.scichina.com info.scichina.com
《中国科学》杂志社
SCIENCE CHINA PRESS

面向感知中国的新一代信息技术专刊 · 论文

移动社交信息智能态势感知分析

萧海东[①②*], 陈宁[③]

① 中国科学院上海高等研究院, 上海 201210
② 上海交通大学中美物流研究院, 上海 200030
③ 华东理工大学信息科学与工程学院, 上海 200237
* 通信作者. E-mail: xiaohaidong@gmail.com

收稿日期: 2014–10–31; 接受日期: 2015–04–16; 网络出版日期: 2015–05-07
上海市科学技术委员会课题 (批准号: 12ZR1415200, 14DZ1119100)、国家自然科学基金 (批准号: 61271349)、中国科学院战略性先导科技专项 (批准号: XDA06010800) 和智慧城市综合运营服务平台 (批准号: Y342341E01) 资助项目

摘要 随着微博、社交网站、移动互联网的快速兴起, 网民通过信息网络获取信息、表达诉求和公共参与日益增多, 舆情监测和引导面临的挑战日益严峻. 针对移动社交网络环境下舆情态势感知存在的问题, 本文提出了智能态势感知方法. 智能态势感知结合了层级时序记忆, 提取的信息流特征更适合于不确定性高的无标度信息动态环境, 能够有效降低历史虚警数据的干扰; 基于自学习的鲁棒特征筛选方法, 实现了对提取节点数据特征的自动筛选, 并在部署移动社交网络服务器应用的环境中以及无人值守的业务情景中实现了态势感知数据的自动汇聚; 构建了态势知识库, 使态势特征更加鲁棒, 不因小世界网络拓扑的动态改变而丢失特征或造成虚警, 使态势感知更适用于小世界网络环境; 提出了空检验矩阵, 对态势可视化结果进行精练并剔除错误匹配点, 具有加速态势可视化的特点. 随着移动社交网络环境的发展成熟和网络用户的成倍增长, 智能态势感知为进一步提升应急信息平台的智能化程度、实现有效的舆情监测提供保障.

关键词 移动社交网络 舆情监测 智能分析 态势感知 小世界

1 引言

微博、社交网站、移动互联网的快速兴起, 网民信息传播空前便捷, “无处不在、无时不有” 成为现实. 同时, 微博 [1] 中存在的大量虚假、失真的言论和信息, 使微博无序化、暴力化倾向明显, 影响网民对现实世界的理性判断, 给正常社会秩序带来冲击. 面对移动互联网的壮大而带来的舆情监管问题, 面对移动社交舆情信息传播网络所呈现出的小世界特性和无标度特性, 面对信息在微博平台上的传播方式呈现 “焰火效应” 的特征, 舆情监测和引导面临严峻的挑战. 因此, 加强网络管理, 组织力量开展网络舆情信息的挖掘, 检测互联网上民众对事件的态度倾向, 监测对相关部门造成威胁的负面信息, 过滤网络不良信息等, 成为当前维护社会稳定、保障社会公共安全、构建和谐社会亟待解决的问题. 在新兴的移动社交网络环境中, 多源信息的获取不再像静态网络拓扑环境中那样容易, 对海量移动社交数据和信息进行分析和处理, 对舆情传播实施智能化监测, 这些问题都给传统的舆情传播分析技术带来挑战. 面对移动社交网络所呈现出的小世界性和无标度性, 态势感知在框架模型、数据预处理、量化感知、动态预测等关键技术问题上显示出良好的鲁棒性, 成为现实中解决此类问题的最佳选择.

引用格式: 萧海东, 陈宁. 移动社交信息智能态势感知分析. 中国科学: 信息科学, 2015, 45: 783–795, doi: 10.1360/N112014-00261

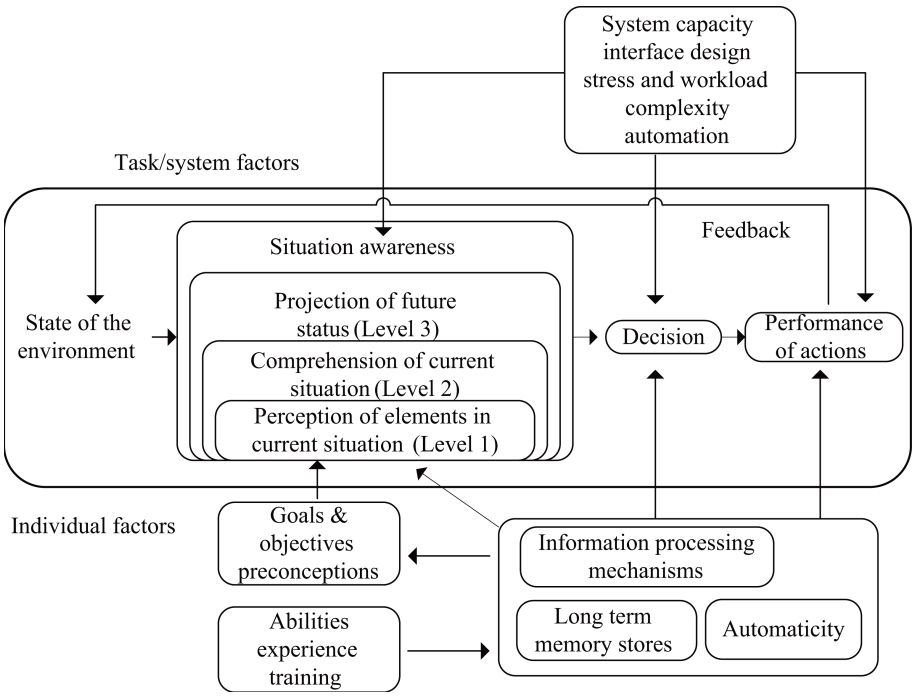

图 1 Endsley 提出的态势模型

Figure 1 Situation model proposed by Endsley

2 现状研究

2.1 态势感知

态势感知 (situation awareness) 概念源于航天飞行的人因 (human factors) 研究 [2], 并在军事战场、核反应控制、空中交通监管 (air traffic control, ATC) 以及医疗应急调度等领域被广泛地研究.

Endsley [2] 提出的比较经典的态势模型, 如图 1 所示. 第 1 层用于感知目前环境的状态、属性以及动态相关因素和环境影响关键因素; 第 2 层通过对关键因素的综合以及显著事件和因素的关联关系, 理解当前态势; 第 3 层反映目前最近趋势变化的情况, 并提供态势预警使应对决策更加科学和及时.

另外, Bass [3] 首次提出了网络态势感知概念, 并对网络态势感知与 ATC 态势感知进行了类比, 旨在把 ATC 态势感知的理论和技术借鉴到网络态势感知中去.

2.2 移动社交网络及其舆情预测

移动社交网络概念强调了社交实体之间存在或多或少的关系, 彼此相互依赖而不是相互独立. 在实际应用中, 一个移动社交网络中的实体或关系往往具有不同的特征, 比如类型、强度、稳定性等 [4]. Watts 等 [5] 的研究为移动社交网络的特征研究奠定了基础. 移动社交网络具备典型的小世界特征, 即网络中的绝大多数节点都是互不相邻的, 但是大部分节点之间只需经过少量几步就可以彼此到达. 另一个重要特性 —— 无标度性基于 Barabdsi 和 Albert[6] 提出的无标度网络 BA 模型.

关于网络舆情的研究已逐渐展开, 其预测是建立在网络热点话题发现的基础之上. Allan 等 [7] 一直致力于话题检测和跟踪 (topic detection and tracking, TDT) 项目的研究, 对新闻媒体信息流进行新话题的自动识别和已知话题的持续跟踪. Chen 等 [8] 提出了生命周期理论, 用来建立话题发现和追踪

模型. Chen 等 [9] 侧重于对网络热点话题的研究, 找出热点话题形成的规律. Jamali 等 [10] 通过分析社交网络的特征和用户评论数, 提出了基于分类和回归体系的网络话题流行度预测算法, 并使用网站 Digg 的数据验证该算法的有效性. Song 等 [11] 利用语义分析算法, 预测人们接收和发送电子邮件的行为. Zhang 等 [12] 针对网络论坛中的热点话题, 提出了基于数据挖掘和小波分析的趋势预测算法. 综上所述, 在移动社交网络舆情预测方面, 目前研究者们采用的大多是常规的统计学、概率论、时间序列分析等方法, 虽然有了较多的成果, 但是预测的效果差强人意. 溯其原因, 主要是由于互联网属于一个复杂的巨系统, 具有小世界和无标度的特征, 网络话题在互联网中的传播必然受到互联网结构、网络社区结构、用户群体数量、用户行为、话题内容和人为因素等多方面、多维度的影响. 因此, 传统的分析方法无论在性能或效率上都较难满足网络话题趋势预测所要求的精度. 结合数据融合和态势理论的研究逐渐在移动社交网络领域兴起, 如在网络信息宏观态势分析方面, 主要关注海量网络信息中关键信息点的量化分析与评价、网络信息在时间维度与空间维度上的发展态势分析与预测等. 主要研究内容包括文本内容的涌现态势发现 (emerging trend detection, ETD) 和时序挖掘 (temporal text mining, TTM). 通过研究实际论证发现, 普通网络态势感知技术并不能统一部署到小世界、无标度的实际移动应用信息环境中去, 需要研究一种新的态势感知方法来解决态势感知问题.

3 移动社交网络环境下智能态势感知分析

3.1 技术路线和系统框图

移动社交网络环境下智能态势感知的技术路线和系统如图 2 所示, 数据主要来自移动社交网络中传播的微博信息、社交即时通信软件空间统计信息等, 通过服务器爬虫程序和微博运营商提供的 API 接口完成数据采集. 首先, 从数据中抽取节点用户信息和用户关系如好友、关注量等基本信息; 其次, 基本信息利用小世界特性和无标度网络理论构建网络拓扑, 研究移动社交网络中信息传播的特性, 如话题热度的变化和话题传播信息的空间聚集度等, 利用金融工程和信号处理技术对指标时间序列进行预处理, 完成舆情数据多维分析; 最后, 利用时序层级记忆 (hierarchical temporal memory, HTM) 智能算法对时序舆情态势相关的多维特征信息流进行融合处理, 结合在上阶段多维分析时产生的移动社交网络拓扑结构, 完成舆情态势信息融合和可视化.

3.2 移动社交网络拓扑形成

移动社交网络是一种复杂网络, 具有小世界特性和无标度特性, 是以社会网络为基础的信息传播网络, 是由多个点 (行动者) 和各个点之间的连线 (行动者之间关系) 组成的集合. 网络节点是信息传播的主体 —— 人, 节点之间的边是信息在两个主体之间的成功传递, 网络中节点关系就是群体内行动者之间的信息交流、认识和信任等关系. 通过服务器 API 接口和爬虫程序采集相关节点信息, 构建移动社交网络拓扑, 其基本原理如下:

(1) 设置初始化参数. m_0(初始网络节点数)、n_0(初始网络边数)、t(演化时间)、r_0(节点的增添速度)、r_1(边的增添速度).

(2) 生成原始的移动社交网络拓扑图. 在初始化参数的基础上, 生成出原始网络的邻接矩阵, 此邻接矩阵是一个对称矩阵, 然后随机生成 m_0 个节点, 根据邻接矩阵确定各个节点之间的连接关系, 画出此网络的拓扑图.

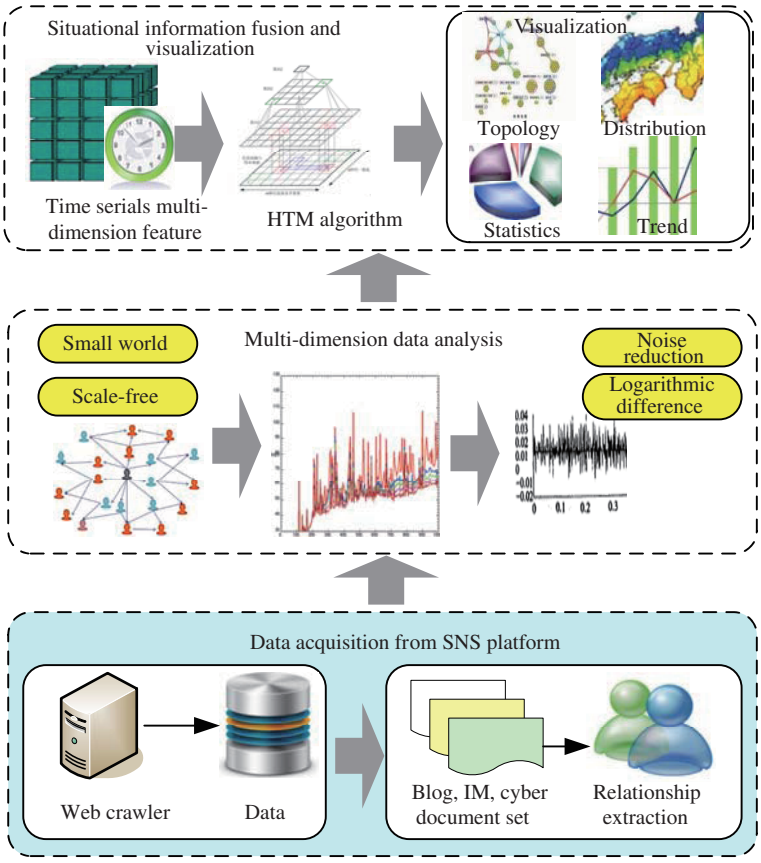

图 2　技术路线和系统总体框图

Figure 2　Technology roadmap and overall system block diagram

(3) 在原始信息传播网络拓扑图的基础上进行 t 步演化: (i) 新增节点. 随机生成一个新的节点, 此新节点所带有的边数 m 也是随机生成的, 在已有的节点中以概率 $\frac{M}{m_0+t}$ 随机选择 M 个节点构成节点集 S, 计算节点的相连数 k, 以概率 $\frac{k_i}{\sum_{j\in S} k_j}$ 在节点集 S 中随机选择 m 个节点与新生成的节点相连接, 则得到增加了新节点的网络拓扑图. (ii) 随机选节点对连接. 随机在当前存在的节点中以 r_0 为比例选择节点对, 在被选中的节点对之间添加边, 将其对应的矩阵中的元素变为 1. (iii) 随机选节点连接. 随机选择一个节点 i, 选择 i 所拥有的好友总数中 $\frac{k_j(k_i-1)}{\sum_{j\in S} k_i(k_j-1)}$ 比例个节点, 在选中的节点中随机选择节点 i 的两个邻接点, 在被选中的邻接点之间添加边, 将其对应的矩阵中的元素变为 1.

(4) 生成最终的网络拓扑图. 经过 t 步演化之后, 按邻接矩阵中反映出来的点与点之间的关系, 连接相应的节点对, 生成最终的移动社交网络拓扑图.

3.3　舆情特征度量指标构建

要对舆情特征进行科学分析, 除了移动社交网络拓扑作为数据分析的平台, 还需要特征度量体系的支持, 这也是实施移动社交网络舆情态势评价的基础. 移动社交网络中信息动态性强、传播迅速, 因此, 特征度量指标的设计既注重整体与重点相结合、可测性和可行性, 又注重移动社交网络舆情生成、扩散、衰退平复的发生周期. 一方面, 它要求评价体系通过各项指标的相互配合进而充分、全面、系统地展现移动社交网络舆情变化, 同时保证各个具体指标在涵义、口径范围、计算方法、计算时间和空

表 1 移动社交网络舆情评测基础指标体系

Table 1 Public opinion evaluation bases of mobile social network

Level 1	Level 2	Level 3
The dynamic change of public opinion	Information variation	Information summary under topic per day
		Rate of topic increase
	Indicator of time effectiveness	Release time of first message
		Update time of last message
		Prescription parameters
	Signature measurement	Signature information
	Publisher effection	Effection measurement
	Diffusion	Time serials of information browse and
	Network distribution measurement	variation of network traffic distribution feature
Public opinion mining	Attention measurement	Publish information variation
		Review variation
		Response variation
	Content sensitivity	Sensitive information filtering results
	Attitude tendency	Tendency analyze result

间范围等方面的一致性. 另一方面, 它要求移动社交网络舆情态势评价指标体系在关注整体性的同时, 必须突出重点反映本质、揭示实质. 建立的基础指标体系如表 1 所示.

在基础指标体系之上, 选择与舆情态势关联度强的指标构建特征, 如话题热度. 定义一个话题 i 的话题热度为第 t 个统计时间内与该话题相关的网络页面的数目、发文量、点击量、转载量、回复量的乘积占该时间段内所有话题统计量乘积之和的比例, 计算公式如下:

$$\text{Topic}_{(i,t)} = \frac{f_i(t) * s_i(t) * g_i(t) * z_i(t) * h_i(t)}{\sum_{j\in\{T_t\}} f_j(t) * s_j(t) * g_j(t) * z_j(t) * h_j(t)},$$

式中, T_t 表示第 t 统计时间内的所有话题, $f_i(t)$ 为第 t 统计时间内与话题 i 相关的网络页面数, $s_i(t)$ 表示在第 t 统计时间内与话题 i 相关的信息在互联网上的累计发文数, $g_i(t)$ 表示在第 t 统计时间内民众点击与话题 i 相关的全部信息的点击总量, $z_i(t)$ 为网民就话题 i 的相关主贴进行转载的总量, $h_i(t)$ 为第 t 统计时间内网民就话题 i 相关的信息跟帖、回复、评论的总数量. 微博平台的信息传播呈辐射状扩散, 在通常情况下, 普通用户发布内容传播有限, 但意见领袖等具有较大影响力的人对信息具有很强的推动和扩散作用, 极大扩展传播范围. 我们可以用信息涉及的用户群体来描述其传播范围:

$$\text{Scale} = f_1|u.\text{Followed}| + f_2\sum_{j=1}^{n}|u.\text{followed}_j.\text{Followed}| + f_3\sum_{j=1}^{n}\sum_{k=1}^{n}|u.\text{followed}_j.\text{followed}_k.\text{Followed}| + \cdots,$$

这里 f_i 表示信息在第 i 次转发时可能的转发概率, u.Followed 表示某用户的所有关注者, u.followed$_j$.Followed 表示用户的关注者 j 的关注者. 由于几层用户转发以后, 涉及的关注者呈指数级增加, 计算量巨大, 完全遵照转发概率 f_i=1 时, 极有可能迅速覆盖全网. 这种情况在实际中没有发生, 正是因为并非每个节点都对信息转发或及时转发, 使转发概率在某些用户节点迅速衰减, 限制了传播范围.

3.4 小世界传播网络信息传播聚集分析

通过移动社交网络拓扑构建和舆情传播特征属性分析, 我们得到影响移动社交网络舆情形成的基

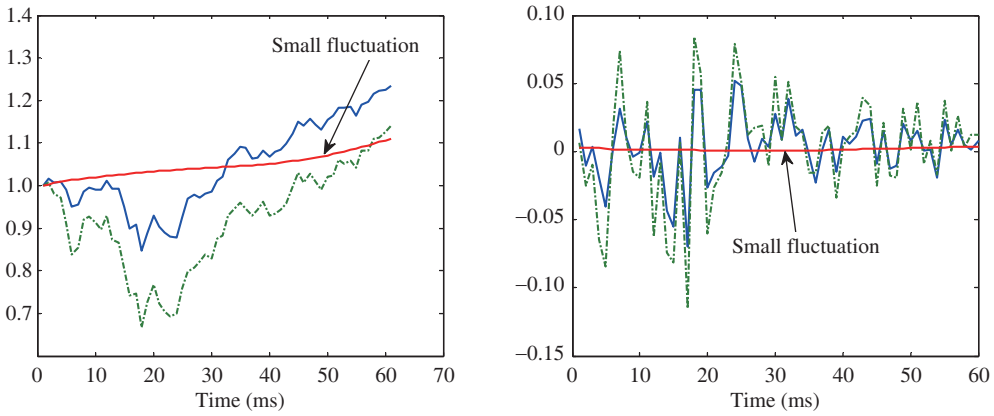

图 3　对数差分数据处理

Figure 3　Logarithmic differential data processing

本数据集, 如何在这些数据集上进一步形成态势表征是需要解决的关键问题. 目前, 大多数舆情研究是在数据层上统计, 在趋势演化层面的研究还较少. 针对移动社交网络中所面临的具体信息环境, 从时间、空间和话题热度、聚集度等多个维度对舆情演化的程度建立测量指标体系, 采集舆情表征数据, 先根据小世界传播网络中信息转发概率 a_i 选定阈值层 E_n 上的活跃边缘信息节点, 通过计算 E_n 内的友元与其周围节点间的关联度大小, 确定两节点间随机捷径连接概率 P, 根据 P 与指定概率 k 的大小关系确定周围节点是否为邻元. 若为邻元, 将该邻元及其友元接受到候选边缘集合 E'_n 中. 将 E'_n 内的候选边缘节点计算平均随机捷径连接距离 D, 如果 D 大于设定值 θ, 则输出社交网络节点边缘检测结果, 否则重新选择阈值层 E_n, 重复上述过程. 具体步骤为:

(1) 指定小世界传播网络中信息聚集概率 k, 门限 θ.

(2) 按概率 a_i 选择初始阈值层 E_n.

(3) 阈值层 E_n 内的友元点 i 与其邻元 j 按计算捷径连接概率 P; 如果$(P \geqslant k) \cap (D_j \geqslant L \times B_i)$, 那么将该邻元作为候选边界点, 接受到 E'_n 内; 否则标记该邻元及其友元为 0.

(4) 所有 E_n 内的标记点都计算完成后, 计算 E'_n 的平均随机捷径连接距离 D: 如果 $D \geqslant \theta$, 那么显示分割结果; 否则返回步骤 (2) 重新计算.

通过上述计算, 可以得到具备时间空间信息的舆情特征 (话题热度随时间变化及信息传播聚集度表征信息的空间扩散), 同理可以得到更多符合舆情指标体系的其他指标数据. 舆情特征信息无量纲化后, 从时间轴观察, 数据形态非常类似金融市场数据, 随时间波动, 这和舆情本身的演化规律是一致的. 借用金融工程理论, 对数据作对数差分, 保留态势内在特征. 这种方法易于发现数据变化态势分布的尖峰, 从统计学角度来说就是指随机变量在均值附近 (即峰顶) 的概率密度值高于正态分布的理论估计值, 从经济学角度来说这是由于价格波动的聚集性造成的, 当市场金融资产价格发生异常剧烈的波动, 并使这种波动在一段时间内持续走高或偏低的话, 就出现了波动聚集, 如果波动聚集在均值附近就出现尖峰现象. 在舆情信息流随时间变化过程中, 尖峰的出现往往预示着舆情变化的拐点或新阶段的开始.

处理后的数据曲线变化如图 3 所示. 图 3 中, 横轴为时间, 纵轴为无量纲化指标度量, 可以看出, 箭头所示线表示的特征随时间变化不大, 其他另外两条曲线波动就较为明显, 有较明显的尖峰特性显现.

态势的主要研究对象是发展趋势性, 为了降低虚警率, 采用数字信号处理方法对态势特征进行降

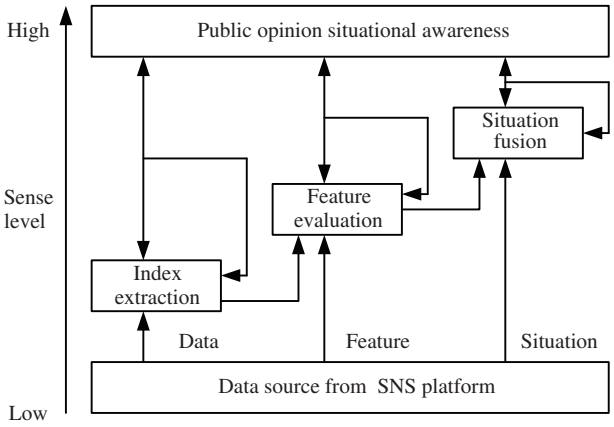

图 4 态势感知框架

Figure 4 Situation awareness framework

噪处理. 一方面, 通过金融数据处理方法, 态势特征信息流已符合一般的信号模式, 如音频信号, 我们可以应用针对音频处理的一些方法来降噪, 同时保证态势变化的主要特征. 另一方面, 为了未来对态势信息的回溯查找, 可以在这个类音频信号流上进行感知 hash, 建立索引, 应用音频内容检索技术快速鲁棒地实现对态势信息的准确定位、回溯. 整个的态势感知框架如图 4 所示, 该框架构建在移动社交应用情景中多维度舆情数据的基础上, 从感知的层次上由低向高划分为指标提取、特征评估、态势汇聚 3 个部分. 其基本原理如下所述.

(1) 首先对数据进行采集, 作为后续处理的对象, 其中数据将包括移动社交网络中的各个方面, 如反应舆情动态变化的信息变化度、时效指标、署名度情况等, 对于涉及移动社交网络环境中舆情挖掘数据可以通过信息关注度变化、评阅量变化、内容敏感度、态度倾向性等方面获得.

(2) 对每个层面的数据信息, 提取时空关联特征, 得到多层面的局部时空对象的特征表述; 然后对这些特征信息进行评估筛选, 摒弃权重弱化的特征, 保留抗数据污染强的鲁棒特征. 在特征评估筛选过程中, 对特征样本进行预设定的攻击, 得到特征集和安全攻击反馈数据集, 一并作为后面层级时序记忆 (HTM) 方法的样本数据.

(3) 基于 HTM 的鲁棒态势汇聚方法, 确定 HTM 网络的分层结构, 将样本数据转化为分层结构感知数据, 并训练它. 结合多层面信息特征表达的互补优势, 基于半监督学习的 HTM 鲁棒态势汇聚器, 完成样本数据在空间阶段和时间阶段的学习, 同时得到 HTM 量化中心数据描述集, 存储到态势知识库并构建推理索引.

(4) 对于态势感知结果, 在指定时间节点或一段时间段进行基于时空检验矩阵运算, 对错误的虚警匹配删除, 精练后的结果将可视化展现.

(5) 同时态势感知的时空检验矩阵参数可作为学习训练各级信息处理层的负反馈输入, 一方面保证态势感知体系在受到突发事件重创时能自行修复, 恢复态势数据的上行通畅, 增加系统整体鲁棒性, 另一方面, 为 HTM 网络提供更多的学习数据, 使 HTM 量化中心数据集更加贴近实际.

该框架有以下特点: (1) 对于数据信息的处理可以在线也可以离线, 虽然在特征提取时会涉及一些计算量较大的操作, 但兼容离线方式, 使得系统响应并不受学习影响; (2) 数据在感知层次上逐级提炼, 在保留信息特征同时并不形成数据的爆炸, 极适应不断增长的大规模数据集的需求. (3) 在数据信息、特征信息、态势信息、特征筛选、可视化处理等方面, 都强调数据的时空关系, 有助于形成态势推

演, 理解舆情信息演化. 该框架涉及到的关键技术有: 基于层级时序记忆 (HTM) 的态势指标汇聚、基于时空检验矩阵的态势可视化.

3.4.1 基于层级时序记忆 (HTM) 的舆情态势指标汇聚

时序层级记忆智能算法是一项对大脑新皮层进行建模的技术. 大脑新皮层占了大约 75% 的人脑容量, 负责所有高层次的理解, 包括视觉、听觉、语言、触觉等. 因为 HTM 是从生物学中得到的, 所以它适合那些人类非常容易做到而对计算机来说非常困难的工作, 例如物体的识别、做出预测、理解语言、在复杂的数据中发现模式等. HTM 是一个记忆系统, 随着时间变化, 它通过感知数据来学习它的世界, 并从数据中抽象出高层的概念. 抽象允许 HTM 网络进行一般化, 并对于传统计算机编程处理的严格规则提供灵活性和效率. 例如, 在不完整或是模糊不清的数据呈现中, 模式能够被学习并识别出来. 通过组合模式学习的记忆与当前的输入, HTM 网络能够预测下一步可能发生什么.

HTM 网络的设计确定了分层结构的大小与架构, 然后为分层结构提供感知数据来训练它. 感知数据来自应用业务中的历史数据. 重要的是在分层中, 有许多数据用来训练, 而且数据具有时间性这一基本元素. 在移动社交网络舆情信息流分析中, 为了进行有效的学习, 都需要在时间的流逝中观察一组模式. 对于一个信息流处理节点, 不管它在分层结构中的位置, 它的输入都是一组模式构成的时间序列. 如归一化处理后, 话题热度为 3, 聚集度 1, 关注度为 1, 则空间矩阵化为 [3, 1, 1]. 在图 5 中标识为 a 的第 1 层节点, 它的输入对应着一个 “拐角形” 特征描述, 如果空间矩阵向右移动一帧, 也就是在下一个时刻, 它的输入对应的是一个变化的 “拐角形”. 这些输入对于一个节点来说, 就是一组模式构成的时间序列. 这个结构由 3 层构成, 数据的输入在最低层, 节点在每个网格里表示, 顶层节点用来实现最终的态势汇聚. 中间分层节点数指数级扩展, 可以有效实现大规模信息流的态势汇聚. 图 5 中输入的特征矩阵为 3×3 大小, 每 4 个下层特征描述区域与上层一个节点对应, 如对于第 1 层的 a 和 b 节点, 下层分别被标记为 A 和 B 的特征区域对应, 同样的道理, 第 2 层的 c 和 d 节点与最下层的 C 和 D 特征区域对应. 不同信息流特征描述在经过这个节点的可接受区域时观察得到一些模式. 把这些模式进行分组, 那些属于同一事物的变体的模式属于同一组. 变体的来源之一是特征与舆情观测标准的相对偏离, 另外一个就是随机噪声. 加入一个节点能够将同源变体的模式分在一组, 那么该组就是该事物变体的恒定体, 被认为是同一个特征的汇聚. 一旦形成分组, 节点就可以产生输出.

3.4.2 基于自学习的鲁棒特征筛选

首先建立预设的移动社交网络舆情影响要素列表, 并根据类型和影响作用, 设置相应权重, 对舆情波动造成的影响、强度越大, 权重越大. 鲁棒特征筛选如图 6 所示. 舆情影响要素在不同的移动社交网络节点有不同的配置, 如用户关注度高和信息更新快的地方 (知名博客、应用门户等) 需要提高对关注度和话题热度的权重, 而对好友多、转发频繁的节点则要重点考虑话题扩散因素. 舆情演化的要素来自移动社交网络信息传播环境, 通过对舆情影响值和阈值的对比判决来提取基本特征集, 然后进入特征空间自学习阶段, 对节点接受到的模式通过最近邻比率匹配原理进行鲁棒特征筛选. 对于每个输入的模式, 计算这个模式与已经存在的量化中心之间的欧氏 (Euclidean) 距离的值, 如果该值在一个数值 D 的范围内, 说明已经有这样的量化中心, 不用增加; 否则向特征空间学习池中增加这个输入模式, 即一个新的量化中心. 数值 D 的大小影响量化中心的数量, 如果过大, 就会把一些不相关的模式组合起来; 如果过小, 就会产生过多的量化中心.

一般来说, 在开始的学习阶段节点都会快速地形成一些量化中心, 随着时间的推移, 单位时间内

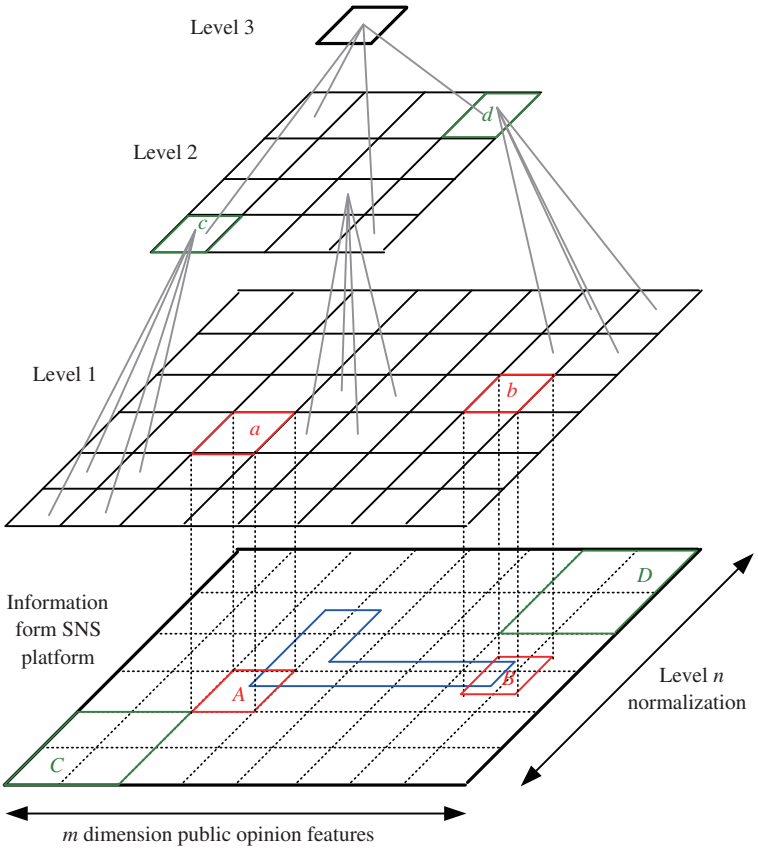

图 5 基于层级时序记忆 (HTM) 的态势指标汇聚网络结构

Figure 5 Indicators converged network architecture based on hierarchical timing memory (HTM)

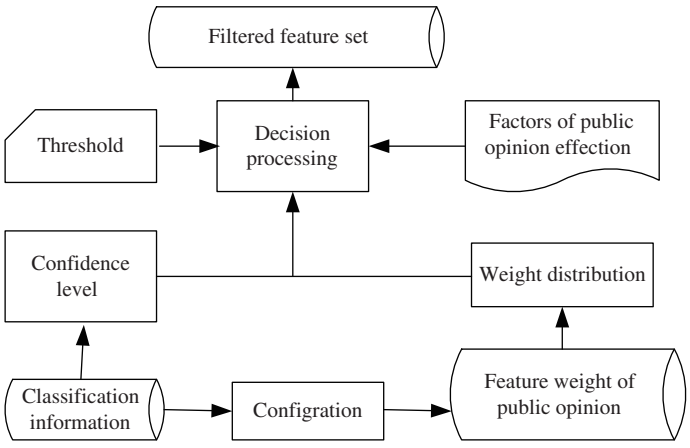

图 6 鲁棒特征筛选

Figure 6 Robust feature selection

增加的量化中心数量越来越少, 直至某个时间期间内小于一个阈值则停止. 鲁棒特征筛选的目的在于增加描述舆情信息特征的鲁棒性, 消除由噪声产生的一些模式.

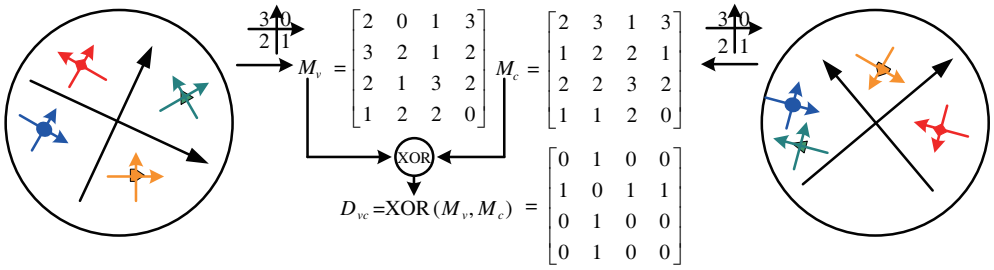

图 7　可视化匹配特征精炼示意图

Figure 7　Refined schematic of visualization matching features

在提高鲁棒性方面, 我们采用以下措施. 在做特征判决时, 将最有价值的信息集中到大特征值对应的特征向量, 避免虚警和噪声带来的误差. 其实质是将信息流的局部特征和全局特征提取方法相结合, 构建以下优化问题:

$$\arg\min_{Y}\sum_{i=1}^{n}\sum_{j=1}^{n}W_{ij}|y_i-y_j|^2+\lambda\sum_{i=1}^{n}\sum_{j=1}^{n}W_{ij}|y_i-y_j|^2,$$

其中, 第 1 项为局部特征距离之和, 其目的是使低维信息流特征数据保持原有的局部结构. 第 2 项是全局特征距离之和, 是正则化项, 其目的是使特征判决时使上述最优化问题求解稳定, 同时避免过拟合. 正则化是通过约束假设空间使安全经验风险最小化, 上式符合 Tikhonov 正则化优化问题, 可以采用标准快速算法精确求解.

3.4.3　基于时空检验矩阵的异构网络环境态势优化、态势可视化

通过对舆情信息态势的感知, 得到移动社交网络局部节点的粗略辅助决策结果, 在信息环境被污染时有可能存在舆情监测偏离匹配的情况, 同时态势数据集不直观. 因此需要对结果进行精炼和优化, 完成态势可视化. 目前的态势数据处理方法是基于时空关系一致的, 如股票价格波动趋势、证券市场波动态势等, 但观测维数太低, 效率不高, 特别是当舆情特征尖峰出现, 短时间内海量转发数据突发时, 所以不能直接应用于移动社交网络环境下的态势可视化. 本文从特征的角度考虑, 提出了新的方法. 基于时空检验矩阵的基本思想是将态势特征点之间的时空相互关系转化为编码, 然后通过编码逻辑运算, 快速检测出特征偏离匹配编码, 并快速剔除. 对于态势的可视化分两步来完成. (1) 精炼匹配出的特征; (2) 精炼态势演化过程.

假定可视化态势片段 V 与态势特征结果集中的时空片段 C 存在匹配特征对: $\{x_i^{(V)}, x_i^{(C)}\}_{i=1}^{n}$, 匹配特征精炼处理见图 7.

具体步骤如下: (1) 每个特征点在提取时将空间、时间和主方向记录下来供可视化使用; (2) 以主方向为起点, 以特征点为中心, 分割空间为 p 个等角度扇形区域; (3) 以特征点所在时空位置为基点划分时域推演区间为前后两个区间来明确历史态势和未来态势关系, 这样时空空间可以划分为 $2p$ 个区间; (4) 按顺序对每个区间设好索引, 建立起该特征和其他特征的时空编码关系. 根据这种可视化映射, 对可视化态势片段 V 和态势特征结果分别定义和计算时空检验矩阵 M_V 和 M_C. 时空检验矩阵的任一元素 m_{ij} 表示态势特征点 x_j 相对于 x_i 的时空关系编码, 即以 x_i 为中心将空间划分为 $2p$ 个区间, x_j 的编码由其所在区间的索引来确定. 图 7 是一个可视化态势片段与态势特征匹配的例子, 每个特征点将空间域划分为 4 个区间, 对应的时空校验矩阵 M_V 和 M_C 分别计算出来, 然后进行异或运算, 得

表 2 运行时间分析

Table 2 Run time analysis

Size of data (MB)	Feature analysis (s)	GPU computing (ms)	Disk I/O (s)	CPU computing (s)	Acceleration
10	0.4530121	27.2121	0.4258	0.725698	11.02076
100	1.0270548	260.1248	0.76693	6.3659	21.52417
200	0.647678	110.698	0.53698	2.771458	20.18535
300	3.0890409	853.2569	2.235784	22.2235	23.4252
500	6.0902077	1366.6597	4.723548	35.2358	22.32615
1024	11.2877639	2594.2639	8.6935	68.31312	22.98132
2048	23.4097698	5188.2698	18.2215	136.22458	22.74421
4096	44.3546838	10352.3258	34.002358	271.1114	22.90394
8192	84.952949	20697.369	64.25558	542.72463	23.11739

到异或矩阵 D_{VC}, 通过分析 D_{VC} 中非零元素所在的行和列, 剔除错误的匹配, 在态势推演时, 这种匹配运算将连续进行, 即加入了时间轴. 同时时空匹配空间也划分为 8 个区间. 当匹配运算很多时, 剔除错误匹配会复杂些, 所以需要找出几个特征点作为参考, 再进行矩阵的检测, 剔除就会加速.

本文提出以可视化态势片段索引为行, 态势特征结果索引为列, 构建二维可视化校验矩阵, 矩阵元素的个数为对应匹配特征的个数, 如图 7 每个特征点将空间划分为 4 个区间, 并从 0~3 编号, 以行特征为基准, 列特征根据其所在区间进行编码, 采用最近邻搜索方式, 剔除掉矩阵行列中最大值元素外的其他所有元素, 保留下正确的匹配, 这样处理有较高的效率, 最后用直方图相似选优算法做出相似性判断, 输出匹配结果.

同时, 这种算法设计非常适合异构计算加速处理. 本文搭建的实验环境如下: (1) CPU, Dual core x86 1.6 GHz 64-bit; Memory, DDR3 Single Channel, 1333/1066 MHz, up to 16 GB; PCI-E, 6x lanes PCI-E 2.0; USB. 3x USB 3.0, 9x USB2.0; SATA, 2x SATA Gen2. (2) GPU, Tesla K20 (存储器带宽 208 GB/s, 存储器容量 5 GB, CUDA 核心数量 2496). (3) 软件条件, Ubuntu 14.04, CUDA-6.5.

选择的态势分析数据集分别选择 10, 100, 200, 300, 500 MB, 和 1, 2, 4, 8 GB 9 种规格, 采用的是二维可视化校验矩阵, 通过表 2 量化对比发现在处理大数据集态势分析时算法加速有明显优势.

4 结语

智能态势感知是智能数据挖掘的方法, 不仅可以用于移动社交网络舆情信息平台监测, 还可以拓展到云安全知识挖掘、网络舆情热点话题发现、分类等领域. 智能态势感知丰富了移动社交网络舆情信息平台评估研究的相关理论与技术, 同时也会促进相关领域学科的发展, 最重要的是, 本文的研究能够为信息科学中态势感知这种大规模复杂数据处理难题提供解决思路, 提高信息处理的性能和安全评估决策效率. 随着移动社交网络环境的发展成熟、网络用户的成倍增长、电子商务的加速发展, 态势感知应用相关的系列产品和服务将会迅猛发展和普及.

参考文献

1 Yi Y G. New Media Blue Book: China New Media Development Report No.3. Beijing: Social Sciences Academic Press, 2012 [尹韵公. 新媒体蓝皮书: 中国新媒体发展报告 No.3 (2012). 北京: 社会科学文献出版社, 2012]

2 Endsley M R. Designing for situation awareness in complex systems. In: Proceedings of the 2nd International Workshop on Symbiosis of Humans, Artifacts and Environment, Kyoto, 2001
3 Bass T. Intrusion systems and multisensor data fusion: creating cyberspace situational awareness. Commun ACM, 2000, 43: 99–105
4 Aggarwal C C. Social Network Data Analytics. New York: Springer, 2011
5 Watts D J, Strogatz S H. Collective dynamics of 'small-world' networks. Nature, 1998, 393: 440–442
6 Barabasi A L, Albert R. Emergence of scaling in random networks. Science, 1999, 286: 509–512
7 Allan J, Carbonell J G, Doddington G, et al. Topic detection and tracking pilot study final report. In: Proceedings of the Broadcast News Transcription and Understanding Workshop. San Francisco: Morgan Kaufmann Publisher Inc., 1998
8 Chen C C, Chen Y T, Chen M C. An aging theory for event life-cycle modeling. IEEE Trans Syst Man Cyber A, 2007, 37: 237–248
9 Chen K Y, Luesukprasert L, Chou S T. Hot topic extraction based on timeline analysis and multi-dimensional sentence modeling. IEEE Trans Knowl Data Eng, 2007, 19: 1016–1025
10 Jamali S, Rangwala H. Digging digg: comment mining, popularity prediction, and social. In: Proceedings of International Conference on Web Information Systems and Mining, Shanghai, 2009. 32–38
11 Song X, Lin C Y, Tseng B L, et al. Modeling and predicting personal information dissemination behavior. In: Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, Chicago, 2005. 479–488
12 Zhang H, Zhao B, Zhong H. Hot trend prediction of network forum topic based on wavelet multi-resolution analysis. Comput Technol Dev, 2009, 19: 76–79

Analysis of intelligent situation awareness of SNS

XIAO HaiDong[1,2*] & CHEN Ning[3]

1 *Shanghai Advanced Research Institute, Chinese Academy of Sciences, Shanghai* 201210*, China*;
2 *Sino-US Globe Logistics Institute, Shanghai Jiaotong University, Shanghai* 200030*, China*;
3 *School of Information Science and Engineering, East China University of Science and Technology, Shanghai* 200237*, China*
*E-mail: xiaohaidong@gmail.com

Abstract With the rapid rise of microblogging, social networking sites, and mobile Internet, users' access to information, public participation, and expression of their demands is increasing significantly. Thus, the challenges faced by public opinion monitoring and guidance are becoming increasingly serious. To address the problems of mobile social network public opinion monitoring, this paper presents an intelligent situation awareness analysis. Intelligent situation awareness combines with the timing of memory hierarchy, extracting information flow characteristics better suited to a highly uncertain, scale-free dynamic information environment, effectively reducing the interference of false alarm history data; is based on robust self-learning screening methods to achieve automatic screening of the node data extraction features and to facilitate the automatic aggregation of situation awareness data in mobile social networking application server environments and unattended scenarios; builds situation knowledge to improve robustness and not lose the characteristics or cause false alarms because of dynamic changes in the network topology of small-world, making situation awareness more suitable for small-world network environments; makes empty inspection matrix scouring the results and eliminating false match points to accelerate visualization. With the maturing of the mobile social networking environment and doubling of the

number of users, intelligent situation awareness provides protection to further enhance the level of intelligence of the emergency information platform and facilitate effective monitoring public opinion.

Keywords mobile social networking, public opinion monitoring, intelligent analysis, situation awareness, small world

XIAO HaiDong was born in 1975. He received a Ph.D. degree in Information and Communication Systems from Shanghai Jiao Tong University, Shanghai in 2008. He was a lecturer at Shanghai Jiao Tong University for eight years. Currently, he is an associate professor at Shanghai Advanced Research Institute (SARI), Chinese Academy of Sciences. His research interests include date fusion, big data, situational awareness, SNS, and AI.

CHEN Ning was born in 1979. She received a Ph.D. degree in Electronic Engineering from Shanghai Jiao Tong University, Shanghai in 2008. Currently, she is an associate professor at the School of Information Science and Engineering, East China University of Science and Technology. Her research interests include signal processing and applications, audio watermarking, audio hashing, time-frequency signal analysis, and music information retrieval.