基于话题检测与聚类的内部舆情监测系统

李忠俊
(西南财经大学经济信息工程学院 温江 611130)

摘 要 针对当前外部舆情系统中响应速度慢、准确度不高等问题,提出基于话题检测与分类的内部舆情监测系统,给出了该系统的组织模型、数据结构与运行流程;采用内外结合的频谱话题检测法来发现当前关注的热点,应用话题聚类预测模型对当前热点话题的可能发展趋势进行预评估,并采取相应措施。实验证明,该系统具有较好的舆情预警能力和较快的响应处理速度。

关键词 舆情,内部监控系统,话题检测,话题聚类,预测
中图法分类号 TP391,TP393 **文献标识码** A

Internal Public Opinions Monitor System Based on Topic Detection and Clustering

LI Zhong-Jun
(School of Economic Information Engineering, Southwestern University of Finance and Economics, Wenjiang 611130, China)

Abstract In order to deal with slow response rate and false alarm rate in external public opinions monitor system, a novel internal system was proposed based on topic detection and clustering. And its organization model, data structure and working flows were described as following. And topic center detection method was utilized to extract the current public hotspots. Then topic clustering and predication model could estimate their trends to monitor internal crisis and so on. Simulation results show that the system can perform better and faster than the traditional system in public opinion alarm.

Keywords Public opinion, Internal monitor system, Topic detection, Topic clustering, Predict

舆情监测系统是对现实热点与焦点问题引发的影响力较大、倾向性性较强的言论和观点进行监视及预测的网络应用系统[1]。此类系统在网络管理中具有重要的现实意义:网络管理人员通过此类系统监测信息及其发展趋势,对网民进行有效引导,并对一些网络舆论危机进行积极化解,从而维护网络和现实社会的稳定与和谐[2,3]。目前,国内外已有若干大型的中心式互联网舆情监测系统面世,如:Buzzlogic公司的Insight系统、Visible Technologies的TruCast和TruView服务、Reputation Defender的客户监控网络、北大方正智思舆情分析应对整体解决方案、中国科学院天玑网络舆情监测系统、乐思网络舆情监测系统等[4-7]。这些系统的主要处理对象是外部网络(Internet)中的广域舆情,其处理流程大致分为3个阶段[8,9]:(1)预案制定:监测中心确定需要监测的目标网站和关键词,并规划相应的处理方案;(2)过程监测:采集、汇总目标网站运营中的相关信息;(3)预警与决策:发现可能发生的网络危机,判断其趋势,修正和调整预案,并实施相应的网络管理活动。在实际应用中,上述系统暴露出一些问题,主要包括[10-12]:监测响应速度慢;误报率较高;预案与处理工作比较宏观和抽象,因而自动化程度较低,需要消耗大量的人力资源完成细节工作。而从现有文献看,话题针对性较强、基于网络内部管理的、自动处理程度较高的舆情监测系统还处于研发空白区。

针对上述问题与研究空白,提出了基于话题检测与分类的内部舆情监测系统。本文第1节对现有系统的问题进行分析,并提出相应的解决方案,且对本系统的组织模型、数据结构与运行流程进行详述;第2节给出本系统中的关键技术与算法;第3节通过实验对本系统的性能进行分析;最后是全文总结。

1 方案与模型

1.1 问题分析与解决方案

当前的网络舆情监测系统主要运行于监测点(通常为网络系统,如网站或网络数据库)之外;这些监测系统通过外部接口对多个监测点内的舆情信息进行抽取、识别与筛选,从而实现监测目的;而监测点并不负责信息处理。这种体系结构带来了两个问题:首先是由于舆情监测系统通过网络和外部接口访问监测点,使得监测响应速度较慢;其次是由于监测系统的监测点较多,监测内容较为宏观,对于各个网络系统难以提供个性化和专门化的服务。针对上述问题,本系统将舆情监测系统下放至各监测点中,由监测点进行微观监测跟踪,并向中心监测系统提供报告,这种模式较为个性化,而且能够在一定程度上实现负载均衡。具体实现参见"系统结构与模型"

到稿日期:2012-02-24 返修日期:2012-06-26 本文受教育部人文社科项目(10YJCZH169),四川省金融智能与金融工程重点实验室项目(FIFE2010-P05),西南财经大学校管课题(2010XG068)资助。

李忠俊(1974—)男,博士生,讲师,主要研究方向为管理信息系统、电子商务等,E-mail:lizj_t@swufe.edu.cn。

部分。

当前的网络舆情监测系统采用中心监测模式，尽管可以将众多监测点中的检测数据融合起来，从而实现广域监测，但由于它们的监测流程主要采用中心服务器“关键词”定时轮询法，导致了一些新出现的舆情信息难以及时上报和汇总。此外，中心监测模式主要依靠“既成事实”的言论进行监测和预测。因此，目前的监测系统需要经过较长的监测周期，在拥有较多的数据积累后危机往往已经进入了新阶段，此时才进行管理干预，消耗的相关资源（时间、人力、财力、技术力量）较多。针对上述问题，本系统除采用分布式的监测方案外，进一步采用基于功率谱估计的热点话题发现技术，及时监测新的舆情发生点，并通过聚类预测对话题的发展进行预测，并完成舆情预处理与相关干预。具体实现参见“关键技术”部分。

1.2 系统结构与模型

该系统的结构与组织模型如图 1 所示。

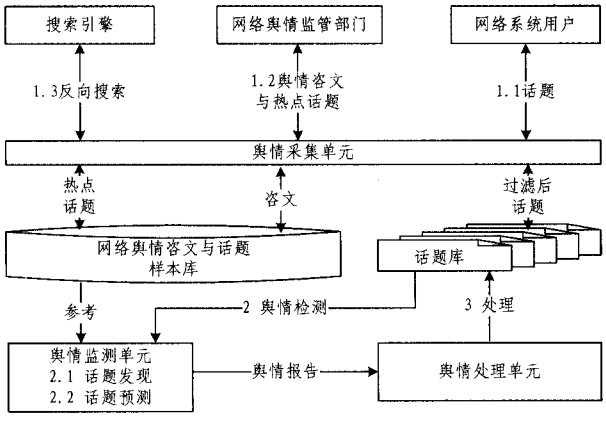

图 1 系统组织结构与运行流程

（1）舆情采集单元：该单元的数据来源包括 3 种。

首先是通过内部信息进行系统内话题监测，主要是对系统内部新近发生的话题进行检测，以发现新的舆情发生点，并作关键词过滤；此外，进行定时扫描，通过聚类预测技术对新生话题的可能发展趋势进行预测，并根据设置进行预案处理。例如：在预处理中，将包含过量敏感词的话题加入高危监控列表以及通知网络系统管理员等。

其次是通过客户-服务器模式获取舆情监管部门的舆情控制咨文。由于话题的扩散通常是受时间和空间限制的，而舆情的形成也是由局部开始，因此监测点的舆情监测必须全面、完整和及时。本系统以监测点为客户端、以舆情监管部门为服务器端进行局部热点话题的汇总融合。即监测点一方面下载监管部门服务器中的舆情控制咨文（主要是热点话题及其特征、预处理关键词等信息），另一方面将自身采集到的同类信息上传，以供其他局部监测点备用。

最后，本系统定时通过互联网中若干重要的搜索引擎进行反向检索，主要根据当前系统内热门话题的关键词，通过搜索引擎的接口获取符合当前舆情趋势的相关网页快照（百度和谷歌等搜索引擎的“网页快照”），从中抽取热点话题及其特征，跟踪与本系统相关的舆情变化情况，并丰富本系统的控制咨文库和话题样本库。

（2）舆情监测单元：该单元主要负责两部分工作。

首先是通过热点话题的检测，发现本系统内的舆情动向与相关线索；定时对系统内话题的主题更新变化频率、回复高低频率比例、有效信息增长速度、话题的关注度等监测数据进行更新与分析，采用功率谱估计技术处理得出该话题的反响程度，并予以排序和记录。

其次是通过热点话题的预测，对本系统内的热点话题进行聚类，通过专家经验库预测其发展趋势；并通过聚类归属，识别非法或垃圾话题。

（3）舆情处理单元：该单元主要对从两个方向上对两个方面的话题进行处理。

首先是对于正面话题，本系统从正方向（SEO，搜索引擎优化）上对其进行优化，主要包括：改变其在网页树中的层次位置（例如顶置、加精、高亮等），冲淡或者提高话题中的关键词密度等。

其次是对于负面话题，本系统主要从反方向（监管与控制）上对其进行处理，主要包括：话题位置调整（例如沉底、推迟发布、移位等）、话题对象处理（例如禁止回复、自动回复或删除话题等）、舆情警报处理（例如报告网络管理员、向舆情监管部门服务器报警、自动截留保存相关话题证据等）。

1.3 系统运行流程

系统的运行流程如图 1 中标识所示。

（1）舆情采集：（1.1）当新生话题进入系统时，进行关键词过滤等处理，采集其特征，注入话题特征库备用；（1.2）从舆情监管部门的服务器中获取控制咨文，用以丰富话题特征库，以备舆情监测和处理使用；（1.3）对搜索引擎进行反向检索，查证话题，并采集当前反向强烈的话题及其特征充实数据库。

（2）舆情监测：（2.1）系统对（1.1）中采集到的话题及其特征，根据（1.2）和（1.3）中采集到的咨文与热点，采用功率谱估计技术进行本系统内的热点话题发现；（2.2）进一步，通过热点话题聚类预测，分析热点话题的发展趋势等变化情况。

（3）舆情处理：根据热点话题聚类预测生成的舆情报告（正负两方面的热点话题列表），进行相关的舆情处理工作。

2 关键技术

2.1 热点话题发现技术

基于功率谱估计的热点话题发现技术是通过既有话题的相关性，估计出接受到的信息的功率（影响程度）随频率的变化关系。功率谱估计主要用于研究信息在频域中的各种变化特征，其目的是根据有限的数据，在相关频域内提取噪声中的变化信息，因此非常适合于对话题的发展趋势进行估计和预测。由于功率谱是由特定信息的各个频率分量融合的功率，因此各个频率分量的功率为特定话题信息中不同频率成分在整个信息变化中所起到的作用。如果某功率谱中的低频部分功率较大，则该信息的变动速率较低；反之，则该信息变动速率较高。而热点话题的发现基于人们对某一个话题的反应。当本站（网站或数据库）中出现新话题时，首先进行第 1.3 节中的预处理，通过基本的过滤、审核和预处理后，进入站内；站内的舆情监测系统定期定时地扫描各个更新过后的话题，并进行第 1.3 节中的分类；并在此时根据功率谱密度估计原理，对分类后监测到的话题的主题更新变化频率、回复高低频率比例、有效信息增长速度、话题的关注度等监测数据进行分析，并通过排序得到热点话题列表。用 $\phi_{xx}(m)$ 表示话题 $x(n)$ 的自相关函数，而 $P_{xx}(\omega)$ 表示它的功率谱密度，则有：

$$P_{xx}(\omega)=\sum_{m=-\infty}^{\infty}\phi_{xx}(m)e^{-j\omega m}$$

而其中：

$$\phi_{xx}(m)=\lim_{N\to\infty}\frac{1}{2N+1}\sum_{n=-N}^{N}x(n)x^*(n+m)$$

代入后可得：

$$P_{xx}(\omega)=\lim_{N\to\infty}\frac{1}{2N+1}\left[\sum_{n=-N}^{N}x(n)e^{-j\omega n}\right]\cdot\left[\sum_{m=-\infty}^{\infty}x^*(n+m)e^{j\omega(n+m)}\right]$$

并将该式求平均，可得：

$$P_{xx}(\omega)=\lim_{N\to\infty}E\left[\frac{1}{2N+1}\left|\sum_{n=-N}^{N}x(n)e^{j\omega n}\right|^2\right]$$

实际应用中，由于网络舆情监测中得到的话题检测数据只能是它的一个数据序列的片段，因此必须采用有限数据序列来估计该话题的功率谱密度。设话题的有限数据序列为：$x(0),x(1),x(2),\cdots,x(N)$，则有：

$$\hat{\phi}_{xx}(m)=\frac{1}{N}\sum_{n=-\infty}^{\infty}x_N(n)x_N(n+m)$$

$$\hat{P}_{xx}(\omega)=\sum_{m}\hat{\varphi}_{xx}(m)e^{-j\omega m}$$

就其分量而言，以主题更新变化频率为例，当某个话题的主题遭到强烈反对或赞同时，将会有大量的舆情导向相反相悖或相同相似的信息注入该话题，而这两个方向上的更新频率的绝对值之和反映了该话题带来的反响情况（回复往往被大量简单的无主题内容充斥，因此仅能作为话题受关注度的参数之一）。

2.2 热点话题预测技术

在实际工作中，新的话题频发。如果对所有话题进行全程跟踪监控，将带来较大的系统开销，因此本系统对新话题采用了基于聚类预测的热点话题预测技术。

聚类预测依据既有的话题样本库进行，通过话题样本及其相关属性，构建了模糊相似矩阵 R[9]（其中行为话题类，列为话题类的各属性）。该类型矩阵通常仅满足自反性和对称性（其中，自反性保证任一个新的话题不能同时属于两个话题类型）。而传递性又是话题预测的必要条件，即假设话题 1 与话题 2 同类，话题 3 与话题 2 也同类，则话题 3 与话题 1 同类。为实现该矩阵的传递性，本系统对矩阵进行了取 λ 截矩阵改造。

(1)由管理员根据经验取 λ 值，模糊相似矩阵中小于 λ 的元素均置为 0，其余的不变。

(2)以行为处理单位，将每一行中的非零元素赋值为行号，构建新的聚类矩阵 R'。从第 0 列开始，执行(3)。

(3)从第 j 列的非零元素 $R'[i,j]$开始，查找与其它列对应的第 i 行的、与 $R'[i,j]$相等的非零元素。如找到，则将该列向量累加到列向量 $R'[j]$中，之后将该列向量清零，如此循环，直到 i 被遍历结束。至此，将获得一个聚类的新向量。列号 j 累进 1，直到 j 被遍历结束，生成新的矩阵。

(4)至此，新矩阵 R' 的每一列均对应唯一的聚类模式。通过查找该矩阵中非零元素的行号，可以找到对应的话题集合样本；而通过比较，可以获得每一个话题聚类中的样本。

(5)当需要对新话题进行估计时，首先根据话题内容及有效回复进行分词，抽取其相关特征，构成话题特征向量；然后计算新话题向量与各个话题聚类的相似程度，以判断其归属，进而通过专家经验库，从话题聚类中观察到此类话题的可能发展趋势。本轮计算中，当前新话题 X 向量的相似度计算公式为：$\mathrm{Sim}=[X*AVG_i+(1-X\times AVG_i)]$。其中，$*$ 和 $\times$ 为模糊运算中的内、外积运算符；而 AVG_i 表示第 i 个话题聚类中所有话题样本的各属性平均值集合。由于分类话题样本中，部分话题属于不和谐范围，因此，热点话题预测后，如发现类似问题，系统将会对这部分话题进行管制和封闭。该功能对网站的内容健康与和谐起到了良好的保护作用。

3 实验结果与分析

本系统在某政务网站与某大型虚拟社区中进行了联合模拟实验，对用户、网站管理人员（版主、各级管理员）等发放 500 份（两套）关于该系统使用情况的调查问卷，共计收到有效问卷 372。两套问卷主要从用户心理角度（通过对普通用户进行问卷调查）和信息和谐角度（通过对版主和系统管理员进行问卷调查）对本系统的性能进行了测试，结果如表 1 所列。

表 1 舆情监测系统使用情况调查统计表（%）

调查项	使用系统后	使用系统前
网站内容和谐程度	92	62
信息有效程度	87	49
信息新颖性	85	62
内容与网站的相关度	93	57
网站内容管理有效程度	91	42
监测响应速度	86	31
监测准确性	83	40
总体满意度	89	47

调查结果显示了使用本系统后，相关网站和社区点击率明显提高；且表 1 显示用户和管理人员对网站的各项主观性能的满意程度均有提高。经过分析可以认为：该系统能够较好地进行站内舆情监测和预警，并根据站外控制中心提供的相关信息对站内信息进行及时调整和监测控制，从而保证站内信息的和谐与有效。此外，版主与网站管理员的问卷调查也显示：该系统能够提供较为准确的热点话题信息列表，并对不合理内容进行及时监测与控制，且响应速度较高（对不良话题的平均监测和控制响应时间为 3 分 17 秒），能够节省大量的时间和人力资源。

此外，网络系统管理员通常担心网络舆情监测系统对网站的流量、客户容量等性能造成影响，而上述性能最终将由系统所在服务器的硬件资源容量决定。因此本系统在联合模拟实验时，以 15 分钟为间隔，对政府网站和虚拟社区所在服务器的 CPU 使用率（应用本系统前后）进行了监测（硬件配置均为：CPU Xeon E5620 2.4GHz×2；内存 4G），以验证本系统对系统性能的影响程度和资源占用率。从图 2 可以看出，由于采用了先进的算法与体系结构，尽管监测需要消耗一定的资源，但本系统由于占用系统资源较少，将部分处理分散到客户端执行，因此对网站系统性能影响较小，具有较高的可用性。

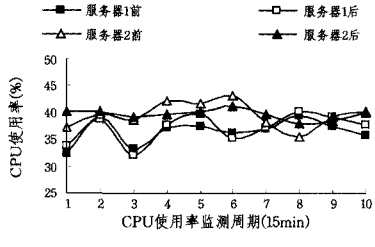

图 2 CPU 使用率监测结果

结束语 基于话题检测与聚类的舆情监测系统，应用功率谱估计热话题检测技术对站内热点进行识别，应用聚类预

测技术对话题进行聚类和预测评估。实验证明，该系统具有较高的响应速度和用户满意度，因而具有一定的应用价值和推广前景。未来的研究包括对等舆情监测系统体系结构、广域舆情监测与控制等。

参 考 文 献

[1] 郑魁，疏学明，袁宏永. 网络舆情热点信息自动发现方法[J]. 计算机工程，2010，36(3)：4-6

[2] 陈莉萍，杜军平. 突发事件热点话题识别系统及关键问题研究[J]. 计算机工程与应用，2011，47(2)：19-22

[3] 宋光慧，聂琰，郭建康. 高校网络舆情监测系统构建及应对机制研究[J]. 计算机与现代化，2011(11)：120-122

[4] 丁菊玲，勒中坚. 基于观点树的网络舆情危机预警方法[J]. 计算机应用研究，2011，28(9)：3501-3505

[5] 丁杰，徐俊刚. IPSMS：一个网络舆情监控系统的设计与实现[J]. 计算机应用与软件，2010，27(4)：188-190

[6] 方薇，何留进，孙凯. 采用元胞自动机的网络舆情传播模型研究[J]. 计算机应用，2010，30(3)：751-755

[7] 闵可锐，赵迎宾，刘昕，等. 互联网话题识别与跟踪系统设计及实现[J]. 计算机工程，2008，34(19)：212-214

[8] 席耀一，林琛，李弼程，等. 基于语义相似度的论坛话题追踪方法[J]. 计算机应用，2011，31(1)：93-95

[9] 龙志祎，程葳. 基于词聚类的热点话题检测算法[J]. 计算机工程与设计，2011，32(6)：2214-2217

[10] 殷风景，肖卫东，葛斌，等. 一种面向网络话题发现的增量文本聚类算法[J]. 计算机应用研究，2011，28(1)：54-57

[11] 洪宇，张宇，范基礼，等. 基于语义域语言模型的中文话题关联检测[J]. 软件学报，2008，19(9)：2265-2275

[12] 吕楠，罗军勇，刘尧，等. 基于话题三层结构模型的话题演化分析算法[J]. 计算机工程，2009，35(23)：71-73

(上接第 213 页)

定义 15(关联规则) 在相容关系下 $a \wedge e \rightarrow x$ 是关联规则，当且仅当 $(N_{T_1}(a) \subseteq N_{T_3}(x)) \wedge (N_{T_2}(e) \subseteq N_{T_3}(x))$。

相应地，定义 15 也可用粒矩阵进行表示。请注意，点的连续性是相容关系下关联规则成立的必要条件。

在表 1 中，条件属性 $C=\{T_1, T_2\}$，决策属性 $D=\{T_3\}$。由表 1 可看到决策规则：

$a \wedge e \rightarrow x; a \wedge f \rightarrow x; b \wedge g \rightarrow y;$

$c \wedge h \rightarrow z; d \wedge h \rightarrow z$

对于 T_1, T_3，有

$$GrRM = (c_{ij})_{m \times n} = U/E_{T_1} \times (U/E_{T_3})'$$

$$= \begin{bmatrix} 11000 \\ 00100 \\ 00010 \\ 00001 \end{bmatrix} \times \begin{bmatrix} 100 \\ 100 \\ 010 \\ 001 \\ 001 \end{bmatrix} = \begin{bmatrix} 200 \\ 010 \\ 001 \\ 001 \end{bmatrix}$$

由定义 14，$k=1$ 且 $GrS_{T_1} \cap GrS_{T_3} = GrS_{T_1}$，故 T_3 中心依赖于 T_1。同理，对于 $T_2 T_3$，虽然 $E_{T_2} \subseteq E_{T_3}$，但是 $GrS_{T_2} \cap GrS_{T_3} \neq GrS_{T_2}$，$T_2$ 到 T_3 只在 h 点连续，$N_Q(h) \subseteq N_Q(z)$。因此表 1 最终可以得到 2 条关联规则：$c \wedge h \rightarrow z; d \wedge h \rightarrow z$。

结束语 等价关系、相容关系都是特殊的二元关系，都可当作粒化方法，并因此衍生相应的知识处理方法和手段。

2005 年起，笔者从二进制粒矩阵理论的角度讨论了 Rough 集理论的上下近似[13]，并用矩阵系统定义了 Rough 集代数定义系统。本文把基于等价关系的矩阵运算拓展到相容关系，定义了基于粒矩阵的相容关系的上、下近似和知识表达系统，给出了相容关系的关联规则发现算法。算例表明了各种概念和定义的可计算性，这些定义可以进一步实现相容关系的知识约简，还可以拓展到普通二元关系中，相关研究还在进行中。

参 考 文 献

[1] Zadeh L A. Some reflections on soft computing, granular computing and their roles in the conception, design and utilization of information/intelligent systems[J]. Soft Computing, 1998(2): 23-25

[2] Lin T Y. Granular Computing: Practices, Theories, and Future Directions[C]// Encyclopedia of Complexity and Systems Science. 2009: 4339-4355

[3] 张铃，张钹. 问题求解理论及应用-商空间粒度计算理论及应用(第 2 版)[M]. 北京：清华大学出版社，2007

[4] 刘清，孙辉，王洪发. 粒计算研究现状及基于 Rough 逻辑语义的粒计算研究[J]. 计算机学报，2008，31(4)：543-555

[5] 苗夺谦，王国胤，刘清，等. 粒计算：过去、现在与展望[M]. 北京：科学出版社，2007

[6] 王国胤，张清华，胡军. 粒计算研究综述[J]. 智能系统学报，2007，2(6)：8-26

[7] Pawlak Z. Rough sets: Theoretical Aspects of Reasoning About Data[M]. Dordrecht: Kluwer Academic Publishers, 1991

[8] 陈泽华，谢刚，谢珺，等. 基于粒计算的 Rough 集模型[J]. 计算机科学，2009，36(5)：200-203

[9] Chen Z H, Lin T Y. Matrix Theory for Tolerance Granular Computing[C]// Proceedings of IEEE International Conference on Granular Computing. San Jose, USA, Aug 2010: 673-676

[10] Lin T Y. Granular Computing on Binary Relations I: Data Mining and Neighborhood Systems[M]// Skowron A, Polkowski L, eds. Rough Sets In Knowledge Discovery. Physica-Verlag, 1998: 107-121

[11] Lin T Y. Chinese Wall Security Policy Models: Information Flows and Confining Trojan Horses[M]// Vimercati S, Ray I, Ray I, eds. Data and Applications Security XVII: Status and Prospects. Kluwer Academic Publishers, 2003: 275-297

[12] Meng J, Wang P, Wang X K, et al. Rule Induction for Tolerance Relation-based Rough Sets[C]// Proceedings of IEEE Granular Computing. Kaohsiung, China, Nov 2011: 445-450

[13] Chen Z H, Xie G, Yan G W, et al. Application of a Matrix-Based Binary Granular Computing Algorithm in RST[C]// Proceedings of IEEE Granular Computing. Beijing, China, July 2005: 409-412

[14] Chen Z H, Lin T Y, Xie G. Knowledge Approximations and Representations in Binary Granular Computing[C]// Proceedings of IEEE Granular Computing. Hangzhou, China, Aug 2012: 75-79