情报学报 ISSN 1000-0135
第30卷第12期1235-1241，2011年12月

JOURNAL OF THE CHINA SOCIETY FOR SCIENTIFIC AND TECHNICAL INFORMATION ISSN 1000-0135
Vol. 30 No. 12, 1235-1241 December 2011

doi:10.3772/j.issn.1000-0135.2011.12.001

基于链接网络图的互联网舆情话题跟踪方法[1)]

朱恒民[1,2] 苏新宁[2] 张相斌[1]

（1.南京邮电大学产业信息安全与应急管理研究基地，南京 210003；
2.南京大学信息管理系，南京 210093）

摘要 互联网舆情演化具有的衍生性和动态性特点，使得舆情话题的跟踪分析相当复杂。为了及时、准确地跟踪舆情的衍生话题，本文在分析网页间的链接关系与网页内容关联性的基础上，提出了舆情演化的链接网络图概念，以及网络图中节点与舆情话题的相关度计量和更新方法，基于此提出了基于链接网络图的舆情话题跟踪方法。实验结果表明，基于链接网络图的舆情话题跟踪方法能够在保持较高准确率的前提下，显著地提高舆情话题跟踪的召回率，并能够从网页的链接中发掘出与舆情话题相关的网页。

关键词 互联网舆情 话题跟踪 链接网络图 话题相关度

A Topic Tracking Method of Internet Public Opinion Based on Link Network Diagram

Zhu Hengmin[1,2], Su Xinning[2] and Zhang Xiangbin[1]

(1. *Research Center of Industry Information Security and Emergency Management, Nanjing University of Posts&Telecommunications, Nanjing* 210003; 2. *Department of Information Management, Nanjing University, Nanjing* 210093)

Abstract The topic tracking of Internet public opinion becomes rather complex because the evolution of public opinion has derivative and dynamic characteristics. In order to timely and accurately track derivative topic of public opinion, a topic tracking method of internet public opinion based on link network diagram is proposed. The relation between links among web pages and their contents is analyzed firstly in this paper. Based on this, the concept of link network diagram and methods for computing and modifying the degree of association between the node of network diagram and the topic of public opinion are proposed. The result of experiment shows that the topic tracking method based on link network diagram can remarkably improve the recall ratio with slight loss of precision comparing with the method of content computing only, and it can even detect some relevant pages from the links of Web pages.

Keywords internet public opinion, topic tracking, link network diagram, degree of association of the topic

1 引 言

互联网舆情就是民众通过互联网对政府管理以及现实社会中各种现象、问题所表达的政治信念、态度、意见和情绪的总和[1]。它是社情民意中最活跃、最尖锐的一部分，最直接、最快速地反映了社会各个层面的舆情状况与发展态势，对社会产生的影响面和影响力越来越大，受到国家有关部门的高度重视。

收稿日期：2010年12月8日

作者简介：**朱恒民**，男，1974年生，南京大学博士后，副教授，主要研究方向：Web挖掘、数据挖掘、知识管理。E-mail：hengminzhu@163.com。**苏新宁**，男，1955年生，教授，博士生导师，主要研究方向：信息智能处理与检索、信息分析与科学评价。**张相斌**，男，1961年生，教授，主要研究方向：线性规划及其逆优化方法与应用。

1） 资助项目：本论文得到教育部人文社会科学项目（项目编号：10YJC870052）、江苏省社会科学基金项目（项目编号：10TQC009）和国家自然科学基金项目（项目编号：70871061和70972083）资助。

互联网舆情随着时间的推进、网民的持续关注和热烈讨论在不断地演化着。与传统媒体的“就事论事”不同，网络传播者泛化以及网络本身具有的虚拟性、匿名性、发散性、渗透性和随意性等特点，使得互联网舆情在发展过程中可能朝任何一个方向发展，路径不确定并经常进行转换，这导致原有的舆情可以衍生出多个与之相关的话题。因此，衍生性是互联网舆情演化的主要特点。由于衍生出的新话题与原有舆情话题在内容上产生了较大的偏移，衍生话题的有效探测是舆情话题跟踪的关键问题，具有极大的挑战性。

动态性是互联网舆情演化的又一特点。在网络的动态信息流中，随着时间的推进，短时间内将产生大量的舆情话题报道，舆情话题关注的焦点在不断地变化。互联网舆情发展的动态性要求舆情话题的跟踪必须是及时和动态地。

网页是一种特殊的文本，页面中嵌入了多个超链接，多个页面之间的链接指向关系在一定程度上反映出页面主题之间的相关性。本文在分析网页间的链接关系与网页内容关联性的基础上，构建舆情演化的链接网络图，提出基于链接网络图的舆情话题跟踪方法，以实现舆情衍生话题的有效探测和舆情话题的及时、准确地跟踪。

2 文献综述

“网络舆论”的概念于2003年年初首次提出。舆论研究是一个新的社会科学与自然科学交叉的研究领域，国内目前在这个领域取得的研究成果相对较少，研究深度也尚待加强。1996年美国国防高级研究计划署(DARPA)提出的话题检测与跟踪技术(topic detection and tracking, TDT)所取得的研究成果极大地推动了网络舆情挖掘和分析技术的发展。与舆情话题跟踪相关的研究工作主要有：Makkonen首次明确讨论了事件演化的概念，指出一个事件可能演化发展成几个相关但独立的话题[2]。国外一些学者通过在语言模型中嵌入时间变量来研究话题的演变：概率时间序列模型dDTM是将时间离散成若干个区间段，对每个区间段的文档采用静态LDA模型进行演化分析。这种方法的计算复杂性将随着时间粒度的细化而迅速增大[3]。TOT模型、DMM和cDTM模型是把时间作为一个连续变量进行演化分析，但是它们把话题本身当作一个常量，考虑时间信息只是更好地分析现有的话题，而不能发现出衍生的新话题[4~6]。此外，Nallapati等指出话题是由事件(event)和故事(story)及其间的联系组成的，采用有向图来描述话题内部的各个事件，主要根据时间的先后来确定事件之间的依赖关系[7]；赵华等提出了一种双质心话题模型，将话题表示成初始质心和当前质心，随着文档的到来而不断更新，体现出话题动态演化的思想[8]；事件框架思想也被用于描述话题的演变，林鸿飞等将话题各个内容侧面定义为“槽”，在进行事件跟踪时，独立计算各个槽的相似度，并通过内容槽扩展的方法解决话题漂移问题[9]。

上述研究工作是通过对语言模型、事件框架等传统的话题模型进行改进，或设计出有向图、双质心等话题模型，以捕捉到舆情话题的动态变化。但是这些模型只是捕捉现有话题的变化，而不能有效地探测舆情衍生出的新话题特征。由于衍生出的新话题与原有舆情话题在内容上产生了较大的偏移，基于内容计算的话题挖掘方法并不能有效地跟踪舆情衍生出的新话题。

链接分析法为话题跟踪中的衍生话题探测提供了新的思路。宋丹等利用网页中的链接信息，通过为种子报道或已确定的相关报道页面中链接所指向的网页加分来更新网页相似度，实现Web话题的跟踪[10]。但是，网页中的链接一般是指向此前已创建的网页，为链接所指向的网页加分是不能及时地评价和跟踪舆情事件的新报道。而且，在舆情事件发生的第一时间产生的种子报道中的链接所指向的页面往往是与舆情事件无关的报道。因此，如何在认清链接动机、链接类型的基础上，充分利用链接信息探测舆情衍生话题，实现舆情话题的动态跟踪，仍需要进一步地研究。

3 互联网舆情演化的链接网络图

超链接是互联网的重要元素，它将分散在各个物理地域的信息有机地结合在一起，使人们能够在网上方便、自在地游历，获取所需的信息。网页间的超链接一方面是引导网页浏览的过程，另一方面也反映了网页创建者的一种判断，即有理由认为，如果网页A存在一条超链接指向网页B，那么网页A的作者是认为网页B包含了有价值的信息。因此，互联网中相互链接的网页之间必然存在着某种关联。本节内容在认清链接动机的基础上，对不同类型、不同结构的链接关系能否揭示出话题之间的关联性进

行分析,基于此构建互联网舆情演化的链接网络图。

3.1 网页间链接与话题的关联性分析

不同动机的链接类型分析。根据网站之间建立链接的动机不同,刘雁书等将网页之间的链接分为推荐链接、合作链接、相关链接、资源链接、通信链接和广告链接六种类型,并分析得出,除了通信链接和广告链接以外,其他类型的链接能够反映出网页话题之间存在的关联性[11]。

不同结构的链接关系分析。根据网页间的链接结构不同,可将网页之间的链接关系区分为直接链接、间接链接、互链接、同被链接和同引链接,分别如图 1 所示。Desikan 等通过分析指出上述五种链接关系往往表明或暗示着网页 A、B 的话题之间是存在关联的[12]。

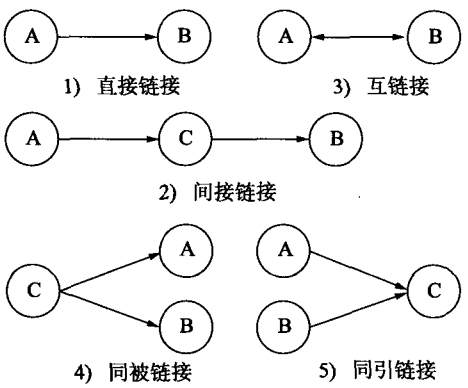

图 1 网页之间多样的链接关系

3.2 舆情演化的链接网络图定义

在互联网舆情事件的产生及其演化过程中,伴随着大量的、持续的报道网页,所有关于舆情话题的报道网页构成了一个文集,文集及其间的链接关系构成了舆情演化的链接网络图。

定义 1 舆情演化的链接网络图 $G(P,E,A)$:P 为网络图中节点的集合,每个节点对应一篇舆情报道网页;E 为节点间的有向连接弧的集合,且 $E=\{<P_i,P_j> \mid P_i,P_j$ 分别为链接网络图中的两个节点,且 P_i 对应的网页存在一条链接指向 P_j 对应的网页$\}$;A 为节点的状态集合,包含节点对应的网页创建时间、网页的输出链接数量,以及与舆情话题的相关度等属性。舆情演化的链接网络图如图 2 所示。

将链接网络图应用于舆情话题的跟踪是基于假设 1 的。

假设 1 在链接网络图 G 中,如果某一网页 A 链接指向一个舆情事件报道 B(或被 B 所链接指向),则网页 A 也可能是该舆情事件的相关报道;如果网页 A 链接指向舆情事件的多个相关报道(或被舆情事件的多个相关报道所链接指向),则网页 A 是该舆情事件相关报道的可能性就较大。

从 3.1 小节的链接动机与链接结构的分析可知,如果在构建链接网络图时不考虑通讯链接和广告链接两种类型,假设 1 是能够成立的。

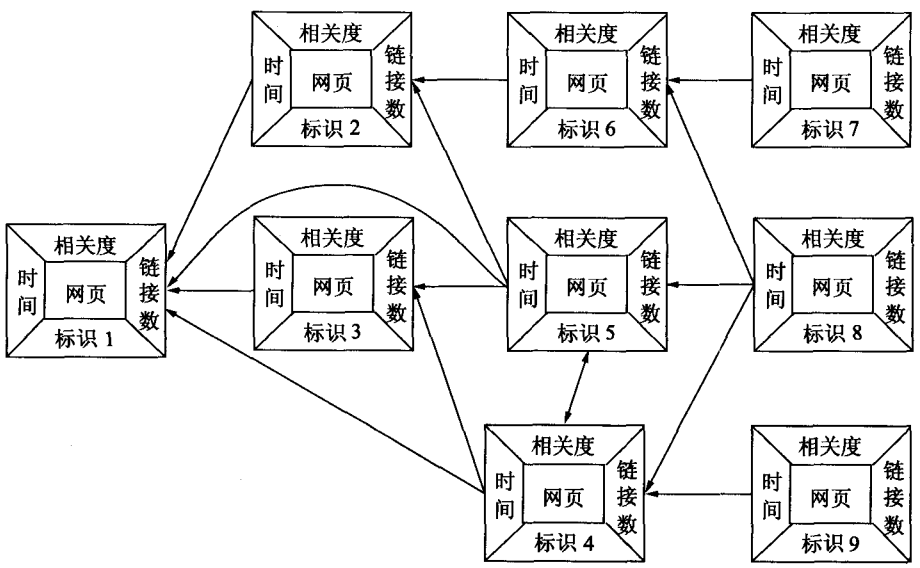

图 2 舆情演化的链接网络图示例

3.3 链接网络图中节点的话题相关度计量方法

网页与舆情种子报道的内容相似度是判断网页与舆情话题是否相关的重要依据。由于网民的泛化和网络本身具有的发散性与随意性等特点,舆情在发展过程中易出现多个衍生话题。例如,本文实验搜集到的2010年"7.28南京爆炸案"舆情事件相关报道中,大约37%的网络报道是关于舆情的衍生话题,如安全隐患和环境风险讨论、舆论应对策略讨论、理赔和善后处理、市政府拟建纪念馆等诸多话题。对于这些内容已经发生偏移的舆情衍生话题的有效识别,需要利用链接网络图中的链接关系进行判断。因此,针对互联网舆情报道这类语料,链接网络图中节点与舆情话题的相关度需要综合链接分析和内容相似度来计量,节点的话题相关度可表达成公式(1)。

$$R = R_L \oplus R_C \tag{1}$$

式中,R_L 为节点的链接相关度,即通过链接分析得到的网络节点与舆情话题的相关度;R_C 为节点的内容相似度,即节点与舆情种子报道的内容相似度;$\oplus$ 为广义加号运算符,考虑到相关度的定义域,$\oplus$ 的选取需满足 $\max(R_L, R_C) \leqslant R_L \oplus R_C \leqslant \min(1, R_L + s \cdot R_C)$,其中 max 和 min 分别为最大和最小函数,s 为调节 R_L 和 R_C 相对权重的系数。

R_C 的计算基于传统的向量空间模型进行,方法已成熟,本文不再赘述。

R_L 的计算。假设 G 是关于舆情话题 T 的链接网络图,如果 G 中一节点 A 通过 n 个链接分别指向了 G 中的其他节点 $P_1, P_2, \cdots, P_n$,取节点 A 与舆情话题 T 的链接相关度 $R_L(A)$ 为:

$$R_L(A) = (R_C(P_1) + R_C(P_2) + \cdots + R_C(P_n))/N(A) \tag{2}$$

其中,$R_C(P_i)$ 为节点 P_i 与舆情话题 T 的内容相似度,$N(A)$ 为节点 A 对应网页的输出链接数量。

3.4 链接网络图中节点的话题相关度调整策略

为了有效地跟踪舆情话题,链接网络图要随着动态的网络信息流新增节点和链接关系(即链接网络图结构)作的动态调整。由假设1可知,链接网络图结构的调整将影响着其他相关节点的话题相关度,因此有必要对节点的话题相关度也进行调整。

链接网络图节点的话题相关度调整策略是:为了保证链接网络图的调整具有较高的时间效率,只对链接网络图中新增节点所链接指向节点的话题相关度进行调整。假设新增节点 A 有一链接指向节点 B,则节点 B 的链接相关度 $R_L(B)$ 依据公式(3)进行调整。

$$R_L(B) = R_L(B) + \frac{R_C(A)}{N(A)} \tag{3}$$

通过上式计算出 $R_L(B)$ 后,再根据公式(1)即可计算出调整后节点 B 的话题相关度 $R(B)$。

4 基于链接网络图的舆情话题跟踪方法

舆情事件一经发生将受到广大网民的持续关注和热烈讨论,短时间内将产生大量的网络相关报道。因此,迫切需要对网络舆情话题的新报道能够及时地跟踪,以捕捉舆情发展的动态变化。

本文基于链接网络图相关理论提出了舆情话题的跟踪方法。该方法是一种增量式话题挖掘方法,链接网络图保存了已读取舆情报道的话题挖掘结果,基于已构建的链接网络图能够计算出新到来舆情报道的话题相关度,而不需要对所有已读取的舆情报道重新开启所有计算。因此该方法是对网络信息流中逐个到来的网页依次话题识别,实现舆情话题的在线跟踪。相对于在整个文集范围内进行聚类的批学习(batch learning)算法[13,14],该方法更能及时、准确地捕捉舆情发展的动态变化。方法的主要步骤如算法1所示。

算法1 TopicTracking_LND( )

Input:网络动态信息流 D,舆情话题 T,种子报道 S;

Output:与舆情话题 T 相关的网页集合;

(1)以种子报道 S 创建节点,初始化舆情话题 T 的链接网络图 G;

(2)for each Web document D_i do {//D_i 为网络信息流中按时间顺序逐个到来的网页;

(3)if ($D_i \notin G$) and ($R_C(D_i) \geqslant \delta_1$) then { //判断 D_i 是否为有待新增且与舆情话题略有相关的网页;

(4)G = AddNode(G, D_i); //将 D_i 添加入网络链接图中;

(5) to extract effective link set L from D_i;//不考虑通讯链接和广告链接;

(6) for each link L_j do {;

(7) if ($P_j \notin G$) then { //P_j 为链接 L_j 所指向的网页;

(8) G = AddNode(G, P_j); } //以 L_j 的锚文本代表页面内容创建节点 P_j,并将其添加入 G 中;

(9) G = AddLink(G, Node(D_i), node(P_j)); } //向 G 中添加从 D_i 指向 P_j 的链接关系;

(10) G = ComputeRela(G, D_i); //计算 D_i 的相关度;

(11) for each link L_j do {;

(12) G = ModifyRela(G, D_i, P_j); } //依据公式3调整 G 中 D_i 链接指向的所有 P_j 节点的相关度;

(13) } G = DelLowRela(G, s,δ_2); //以网络图中新增节点数量达到 s 作为周期,定期删除相关度 $\leqslant\delta_2$ 的节//点,以精简 G 的结构;

(14) } ;

(15) return G 中相关度 $R(D_i)\geqslant\delta_3$ 的所有节点 D_i。

算法1表明,链接网络图是通过依次读取网络信息流中的网页逐步完成动态构建的,步骤(3)至(13)是构建链接网络图的多次循环中关于 D_i 的一次计算过程。其中,步骤(3)至(9)是完成链接网络图节点和链接关系的添加;步骤(10)是计算当前添加节点 D_i 的相关度;步骤(11)至(12)是动态调整 D_i 链接指向的所有节点 P_j 的相关度;步骤(13)是完成图 G 的精简。步骤(15)是返回 G 中相关度大于指定域值的所有节点对应的网页集合。

时间效率是舆情话题跟踪方法考虑的重要指标。算法1中对网络图待添加节点和链接的筛选[见步骤(3)和步骤(5)],以及节点状态的动态调整策略[见步骤(12)]和对链接网络图结构的精简[见步骤(13)],都有助于提高算法1的时间效率。对算法1进行时间复杂性分析。令 n 表示网络动态信息流 D 中的网页数,a 表示每篇文档中抽取出的有效链接的最大数量,则算法1的运行时间为 $O(cn)$,c 为与 a 相关的正常数。因此,算法1的时间复杂性是可行的。

5 实验分析

实验选用2010年"7.28南京爆炸案"舆情事件作为分析对象,实验文集选用2010年7月28日至8月15日期间网易新闻(http://news.163.com)中关于南京塑料厂爆炸案、长沙税务大楼爆炸案、河南商城爆炸杀人案、化工炸弹、环境讨论和建设纪念馆等话题的619篇网页。其中关于南京爆炸案的报道中覆盖了爆炸现场、人员伤亡抢救、政府应对、爆炸原因分析、肇事者调查、理赔和善后处理、爆炸后环境风险讨论、安全隐患排查、市政府拟建纪念馆等主题。实验文集的获取是先通过设置关键词、时间段和来源网站等检索条件,调用通用搜索引擎完成相关话题的检索,然后采用网络爬虫将通用搜索引擎返回的检索结果采集下来。

为了评价基于链接网络图的互联网舆情话题跟踪方法的有效性,实验将该方法同仅仅基于内容计算的方法进行了对比。两种方法及网络爬虫均在Visual C++ 6.0开发环境下程序实现,采用张华平博士提供的中文分词工具ICTCLAS2011(http://hi.baidu.com/drkevinzhang/home)对网页文本内容进行分词处理。实验首先将采集下来的实验文集按网页创建时间的先后顺序进行排序,模拟出网络信息流;分别采用上述两种方法对"7.28南京爆炸案"舆情事件进行追踪实验;结合人工判断对实验结果进行评价分析。

实验采用准确率 P、召回率 R 和综合指标 F 来评价两种方法的话题追踪效果,其中 $P = n_c/n_a$, $R = n_c/n_t$,n_a 表示方法判断出与舆情话题相关的文档数目,n_c 为 n_a 个文档中正确反映该舆情话题的文档数目,n_t 表示由人工判断出实验文集中与舆情话题相关的所有文档数目;$F = P\times R\times 2/(P+R)$。

基于链接网络图的舆情话题跟踪方法要选取合适的话题相关度计算公式[即公式(1)]。考虑到话题相关度的定义域,我们在192条实验数据构成的文集上对比了公式(4)、公式(5)和公式(6)三个话题相关度计算公式对舆情话题跟踪效果的影响,实验结果如表1所示。

$$R = R_L \oplus R_C = \min(R_L + R_C, 1) \quad (4)$$

$$R = R_L \oplus R_C = \min(\sqrt{R_L^2 + R_C^2}, 1) \quad (5)$$

$$R = R_L \oplus R_C = \max(R_L, R_C) \quad (6)$$

从表1可知,尽管公式(4)的准确率比不上公式(5)和公式(6),但是其召回率明显高于它们,而且它的综合指标 F 的平均值是最大的,因此实验将采用公式(4)计算链接网络图中节点的话题相关度。

在实验文集上,基于链接网络图的舆情话题跟踪方法同仅仅基于内容计算方法的对比实验结果见表2。考虑到如果相关度阈值选取的较高,两者的召回率都太小;如果阈值选取的太低,则两者的准确率又比较低,因此文中只列出了相关度阈值在[0.1,0.4]区间的实验数据结果。

表1 三个话题相关度计算公式影响舆情话题跟踪效果的对比实验结果

相关度阈值	P(%)			R(%)			F(%)		
	公式(4)	公式(5)	公式(6)	公式(4)	公式(5)	公式(6)	公式(4)	公式(5)	公式(6)
0.4	100	100	100	40.68	33.90	18.64	57.83	50.63	31.42
0.3	100	100	100	79.66	54.24	52.54	88.68	70.33	68.89
0.2	79.17	100	100	96.61	89.83	79.66	87.02	94.64	88.68
0.1	57.84	62.77	66.67	100	100	98.31	73.29	77.13	79.46
平均值	84.25	90.69	91.67	79.24	69.49	62.29	76.71	73.18	67.11

表2 两种舆情话题跟踪方法的实验结果

相关度阈值	P(%)		R(%)		F(%)	
	内容计算	基于链接网络图	内容计算	基于链接网络图	内容计算	基于链接网络图
0.4	100	100	8.24	59.93	15.23	74.95
0.3	100	100	37.83	83.90	54.89	91.25
0.25	100	99.20	62.92	92.51	77.24	95.74
0.2	100	94.58	87.64	98.13	93.41	96.32
0.15	98.47	88.0	96.25	98.88	97.35	93.12
0.1	89.83	78.76	99.25	100	94.31	88.12
平均值	98.05	93.42	65.36	88.89	72.07	89.92

实验结果反映出,基于链接网络图的舆情话题跟踪方法在召回率上具有明显的优势,这归功于前者具有高效的衍生话题探测能力。例如,当阈值设为0.1时,只有基于链接网络图方法才能跟踪到内容相似度较低的舆情衍生出的两篇报道:“南京‘7·28’纪念馆应该怎样建才合民意?”和“南京拟在爆燃遗址建纪念馆,网友建议责任人出钱”。

表2的实验结果还反映出,基于链接网络图的话题跟踪方法的准确率不及仅仅基于内容计算的方法,这是因为前者话题相关度中的链接相关度分量对话题识别的准确率不及内容相似度,尽管网页间的链接关系与页面内容具有关联性,但是内容相似度仍然是话题识别的最准确依据。

综上所述,基于链接网络图的话题跟踪方法在召回率上具有明显的优势,但对话题识别的准确率不及仅仅基于内容计算的方法。这导致在不同相关度阈值下,两种方法在综合指标F上的表现优势是不同的:在相关度阈值大的情况下,两种方法的准确率都比较高,此时召回率是评价方法表现的敏感指标,因此基于链接网络图的方法在F上表现占优;在相关度阈值小的情况下,大量低相关度的网页通过了筛选,两种方法的召回率都比较高,此时准确率是评价方法表现的敏感指标,因此仅仅基于内容计算的方法在F上表现占优。

评价两种方法的话题跟踪效果需要综合考虑具体应用领域需要和实验表现。从方法的应用领域需要来看,舆情衍生话题的有效探测是舆情话题跟踪的关键难题,具有极大的挑战性。基于链接网络图的舆情话题跟踪方法具有高效的衍生话题探测能力,因此该方法更适合舆情话题的跟踪。从实验结果来看,在不同相关度阈值水平下,两种方法在综合指标F上的表现优势不同,但是从F的平均值来看,基于链接网络图的话题跟踪方法相对较优。

基于链接网络图的舆情话题跟踪方法不仅能够有效地提高在实验文集中跟踪话题的召回率,还可以挖掘出实验文集以外,但有链接指向的舆情相关网页。令N_e表示基于链接网络图方法从实验文集中发掘出的舆情话题相关网页数量,N_k表示基于该方法从链接中发掘出的实验文集以外的舆情相关网页数量,P_k、R_k和F_k分别表示基于该方法在链接中

跟踪舆情话题的准确率、召回率和综合指标。表3列出了不同相关度阈值下基于链接网络图方法从链接中跟踪话题的实验结果。

表3 基于链接网络图方法从链接中跟踪话题的实验结果

相关度阈值	P_k(%)	R_k(%)	F_k(%)	N_k/N_e
0.4	60.76	32.0	41.92	0.24
0.3	69.5	65.33	67.35	0.39
0.25	68.79	79.33	73.68	0.42
0.20	63.64	84.0	72.42	0.42
0.15	54.0	90.0	67.5	0.45
0.1	43.48	93.33	59.32	0.46

表3中 N_k/N_e 具有较大的值表明，实验文集的链接中包含了不少文集以外的舆情相关网页。由于受到软硬件资源的限制，不可能针对整个网络空间中的网页进行话题跟踪，本文提出的基于链接网络图的舆情话题跟踪方法能够从链接中发掘出舆情相关网页，提高了舆情话题跟踪的召回率。与在实验文集中进行话题跟踪相比，基于链接网络图在链接中跟踪话题的准确率有所降低，这是由于后者只是根据链接的锚文本来计算该链接对应节点的内容相似度。一般来说，链接的锚文本往往是网页的标题，虽能点出报道主题，但不能够全面地描述页面内容，因此影响了话题跟踪的准确率。

6 结束语

本文提出的基于链接网络图的互联网舆情话题跟踪方法，充分利用网页间的链接关系与内容之间的关联性，有效解决了内容发生偏移的舆情衍生话题的探测难题，显著提高了舆情话题跟踪的召回率。而且，链接网络图结构和节点状态随着互联网舆情的新报道而动态调整，实现了对互联网舆情的在线跟踪。由于在构建链接网络图时考虑到了网页的有效链接，该方法还能够从网页的链接中发掘出与舆情话题相关的网页。虽然该方法通过考虑网页间的链接关系显著提高了舆情话题跟踪的召回率，但是实验结果反映出该方法在召回率提高的同时准确率略有降低。因此，如何充分利用链接关系的特征提高该方法中链接相关度对话题识别的准确率将是作者下一阶段的主要任务。

参 考 文 献

[1] 姜胜洪. 网络舆情热点的形成与发展、现状及舆论引导[J]. 理论月刊，2008(4)：34-36.

[2] Makkonen J. Investigations on event evolution in TDT [C]//Proceedings of the 2003 conference of the North American chapter of the association for computational linguistics on human language technology. Edmonton, Canada, 2003：43-48.

[3] Blei D M, Lafferty J D. Dynamic topic models [C]// Proceedings of the 23rd international conference on machine learning. Pittsburgh, PA, 2006：113-120.

[4] Wang X, McCallum A. Topics over time: A non-Markov continuous-time model of topical trends[C]//Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining. Philadelphia, USA, 2006：424-433.

[5] Wei X, Sun J, Wang X. Dynamic mixture models for multiple time series [C]//Proceedings of the 20th international Joint conference on artificial intelligence. Hyderabad, India, 2007：2909-2914.

[6] Wang C, Blei D M, Heckerman D. Continuous time dynamic topic models [C]//Proceedings of the 23rd conference on uncertainty in artificial intelligence. Helsinki, Finland, 2008：579-586.

[7] Nallapati R, Feng A, Peng F, et al. Event threading within news topics [C]//Proceedings of the 13th ACM international conference on information and knowledge management. Washington, USA, 2004：446-453.

[8] 赵华，赵铁军，于浩，等. 面向动态演化的话题检测研究[J]. 高技术通讯，2006，16(12)：1230-1235.

[9] 林鸿飞，宋丹，杨志豪. 基于语义框架的话题跟踪方法[C]//中国中文信息学会二十五周年学术会议论文集. 2006：383-392.

[10] 宋丹，林鸿飞，杨志豪. 基于内容计算和链接分析的Web话题跟踪方法[J]. 情报学报，2007，26(4)：555-560.

[11] 刘雁书，方平. 利用链接关系评价网络信息的可行性研究[J]. 情报学报，2002，21(4)：401-406.

[12] Desikan P, Srivastava J, Kumar V, et al. Hyperlink Analysis: Techniques and Applications [R]. Technical report of army high performance computing and research center, 2002.

[13] 王伟，许鑫. 基于聚类的网络舆情热点发现及分析[J]. 现代图书情报技术，2009，18(3)：74-79.

[14] Connell M, Feng A, Kumaran G, et al. UMass at TDT 2004[C]//Proceedings of the 7th topic detection and tracking conference. Gaithersburg, MD, 2004.

（责任编辑 王建平）