万维网 (2014) 17:377–404 DOI 10.1007/
s11280-013-0207-7

基于W2T方法论提取新闻博客热点话题

周二忠·钟宁·李跃峰

收讫日期:2012 年 10 月 8 日 / 修订日期:2013 年 2 月 22 日 /
接受日期:2013 年 2 月 27 日/在线发表日期:2013 年 3 月 26 日

摘要尽管主题检测和跟踪技术已经取得了很大的进步,但大多数研究人员很少关注以下两个方面。首先,主题模型的构建没有考虑不同主题的特点。其次,没有进一步分析决定热点话题形成和发展的因素。为了正确提取新闻博客热点话题,本文基于W2T（智慧物联网)方法论,从新的角度看待上述问题,其中研究了博客用户特征、话题传播情境和信息粒度。统一的方式。首先分析博客用户的动机和特征,了解新闻博客话题的特征。然后将话题传播的上下文分别分解为博客社区、话题网络和观点网络。识别用户行为模式、意见领袖和网络舆论等重要因素来跟踪新闻博客话题的发展趋势。此外,提出了一种博客热点主题检测算法,通过测量持续时间、主题新颖性、用户关注度和主题增长来识别新闻博客热点主题。实验结果表明

E. Zhou · N. Zhong (B)
北京工业大学国际 WIC 研究院,中华人民共和国北京 100124 电子邮件:zhong@maebashi-it.ac.jp

周女士 电子
邮件: zez2008@emails.bjut.edu.cn

钟 N. 前桥工业
大学生命科学与信息学系,460-1 Kamisadori-Cho,Maebashi 371-0816,日本

Y. Li 昆
士兰科技大学科学技术学院,布里斯班 QLD 4001,澳大利亚 电子邮件:y2.li@qut.edu.au

所提出的方法可行、有效。这些结果对于进一步研究博客空间意见领袖的形成机制也具有一定的参考价值。

关键词智慧物联网·信息粒度·主题检测·
意见领袖·话题热度评价

1 简介

博客是一种以网页形式存在的在线日记，作为Web 2.0的产物，提供信息分享和观点互动服务，博客空间由博客和相关链接组成，如今用户越来越依赖这个虚拟的博客空间来满足情感表达、信息检索等方面的需求，博客空间中的主题检测与观点挖掘在信息推荐和商业领域也有着重要的应用价值。新闻博客是一种简洁、及时的新闻媒体，新闻博客中的话题涉及生活的各个方面，来源于社会事件、政治事件、经济事件等新闻事件，吸引了大量网民的关注。尽管话题检测与跟踪技术取得了长足的进步，但研究者们仍然面临着许多问题，首先，信息过载和个性化管理给博客空间中的主题挖掘和观点挖掘带来了很大的问题。

其次，用户对话题热度没有制定统一标准。第三，什么决定博客热点话题的形成和发展仍然是一个尚未解决的问题。

新闻博客热点话题检测与跟踪系统是广义的W2T（智慧物联网)方法论的一个典型用例，该方法论为如何统一人类模型、信息网络和粒度以分析社会世界和网络世界之间的交叉问题提供了一种视角[55, 56]。W2T所指的智慧意味着服务对象不是孤立的，必须考虑相关的上下文。W2T方法论中所指的网络和信息粒度为以人为中心的计算提供了一种手段。在网络的概念上，复杂网络理论强调通过结构变化来分析相关现象或事物。例如，复杂网络理论提出的一些方法被成功地用于发现在线社区或研究社区中的信息传播机制[5, 32]。另一方面，社会网络分析理论强调社区中个体之间的关系，广泛应用于社会学研究。例如，用户角色的识别通常基于社会网络分析理论[59]。至于粒度的思想，粒计算理论提供了一种通过多层次信息粒度解决问题的策略[48]。此外，粒计算是 W2T 方法论的理论背景之一，该方法论用于信息检索和 Web 使用挖掘。例如，强调 Web 信息多种表示形式的信息检索支持系统 (IRSS) 框架就是一个典型应用[47]。在 IRSS 中，文档空间、查询空间、术语空间和检索结果都是粒化的。强调用户行为数据的粒结构的 Web 使用挖掘方法是另一种应用[52]。在该方法中，首先根据用户行为数据的结构对网页中的用户行为数据进行粒化

网站和服务内容,然后推断用户动机和潜在意图。 W2T还强调人类对于各种服务中的应用的影响和特征。如果将W2T相关理论应用到信息推荐系统中,统一人、网络和信息粒度的思想,意味着用户个性、情境和需求是系统的主要输入。这种系统的机制可以解释如下。首先,将上下文映射到网络模型中。然后分析网络的结构和模型内实体之间的关系,并根据网络的特征和用户的个性来分析用户的需求。最后根据用户需求的性质识别出不同的信息粒度,并为用户组织和处理相关信息。

前人研究已认识到用户是网络信息传播中最重要的因素,例如用户思维、交互模式[15]、有影响力用户的参与程度[1,39]、用户动机[29]、用户背景[31]、邻居特征[10]等均会影响信息传播。同时,W2T问题的求解策略更注重人性化。因此,为了发现新闻博客中的热点话题,本文基于W2T方法论来分析博客话题的形成和发展。即首先将话题传播的上下文分解为与用户相关的不同类别的复杂网络。然后在特定信息粒度对应的相关网络中衡量不同类型的用户和网络观点对话题的影响力。最后,本文得出结论:用户动机和行为模式决定了新闻博客话题的突发性和时间性。此外,用户行为模式、网络观点和意见领袖在博客话题的不同阶段都起着至关重要的作用。

本文其余部分安排如下。第2节介绍了热点话题检测的相关工作。第3节分析了相关问题。第4节描述了所提出的方法。第5节通过一些实验测试了所提方法的可行性和有效性。第6节给出了结论和未来的工作。

2 相关工作

在主题检测和跟踪 (TDT) 领域,主题被定义为重大事件或活动,以及所有直接相关的事件和活动[12]。就主题检测技术而言,向量空间模型[12,17,41,43,57] 、概率模型[7,9]和复杂网络理论方法[36,54,58]很流行。他和同事使用增量词频逆文档频率模型和增量聚类算法来检测新事件[17]。

陈等人首先根据时间分布和生命周期提取热门词汇,然后识别关键句,并将关键句分组成代表热门话题的簇[12]。周等人采用基于密度的带噪声空间聚类 (DBSCAN) 方法将词汇分组成词簇,从而提取热门话题[57]。王等人采用使用算术平均数的非加权对组法 (UPGMA) 算法检测新闻话题[41]。王等人提出了一种改进的 k-means

方法来检测热门话题[43]。Brants等利用概率潜在语义分析(PLSA)模型来检测文档中的不同主题[9]。Blei 等利用潜在狄利克雷分配(LDA)方法分析文档中的主题[7]。朱军等根据词语共现构建词语网络，并根据词语网络的小世界结构从文本中提取关键词[58]。石军等根据文本词语的小世界结构分析文本的主题[36]。赵军等根据复杂网络理论和词语的小世界结构从文档中提取关键词。基于向量空间模型的方法因其简单性而被广泛应用[54]。但是向量空间模型缺乏语义相关性，存在特征维数高的问题。PLSA 和 LDA 模型依赖于大量样本数据。基于复杂网络理论的方法比其他方法更复杂。

热点话题是用户广泛讨论并持续关注的话题。他和同事考虑了新闻报道的频率和连续时间来评估话题的热度[17]。龚从用户参与度和媒体报道来衡量话题热度[16]。李及其同事通过结合评论数、评论数、意见数和发表时间来实现博客主题热度评估[24]。对于话题热度评估，上述大多数评估策略主要关注用户的反应，而没有进一步考虑话题热度的本质。即博客话题的热度反映了所有用户的兴趣程度，话题热度的变化意味着用户的兴趣受到某些因素的影响。因此，热度评价不能忽视影响话题发展趋势的因素的变化。

一般来说，当孤立的个体行为发展成群体行为时，往往会出现热点话题。然而，要回答博客话题为何会成为热门话题仍然是一个复杂的问题，因为话题热度与事件、网络媒体、用户等多种因素有关。论文中的研究策略与传统策略不同，因为我们意识到以下现象。首先，用户既重视新闻事件的发展，也关注其他用户发表的观点。因此网络观点被认为是话题增长的重要因素，需要检测该话题的网络观点的变化。其次，网络可以放大意见领袖的影响力，以至于新事件的发展有时是由意见领袖决定的。因此，我们要仔细分析意见领袖的特征和形成条件，衡量意见领袖对话题的影响力。

尽管研究者已经意识到在线社区的演化主要源于用户行为的变化[23]，但他们并没有强调影响用户决策的原因。这也是为什么W2T方法论强调从人的角度理解网络世界中的现象。近年来，许多学者根据W2T的系统框架或方法论重新思考传统问题。在线观点的演化和用户在线行为的分析一直是网络智能应用中的关键问题。例如，刘等人试图将用户观点、产品销售和计算机系统整合为一个实体，这是W2T数据循环系统的直接应用[27]。

Msuical 及其同事采用复杂网络理论的方法，制作了一个

对网络世界中的用户行为和交互进行调查,以提供 W2T 中的个性化服务[30]。

3 问题分析

为了构建有效的主题模型来表示新闻博客主题并合理评估主题,本节研究了传统方法忽略的两个问题。首先,新闻博客和博客用户有什么特点,意见领袖作为特殊的博客用户是如何产生的。其次,如何将主题传播的上下文表示并分解为相关的复杂网络,以及如何识别不同复杂网络中决定新闻博客主题发展趋势的关键因素。

3.1 新闻博客的特征

一个话题的开展和相关博客的特点相关。
一般来说,博客根据帖子内容可以分为三种类型,即专业技术、个人生活和时事话题[34]。专业博客的主题主要是指与专业技术相关的信息,而生活博客的主题则涉及个人生活事务。此外,时态博客中的主题主要与新闻事件相关。

新闻博客继承了时态主题博客的特征。与专业博客话题相比,新闻博客话题往往呈现突发性、时效性的特点。用户出于不同的动机访问不同类型的博客[34]。

例如,用户在专业博客中分享专业知识、相互学习。用户通过个人生活博客构建社交网络或维持社会联系。而博主在新闻博客中发布帖子的目的是揭示新闻事件的真相或引起其他用户对特定事件的关注。来自不同背景的用户为了了解新闻事件的真相或表达自己的个人感受而参与话题互动。话题组中成员之间的关系较弱,因为话题组主要由用户兴趣来维持[40]。当用户对相关主题失去兴趣时,话题组就会解散。因此,用户动机在很大程度上决定了新闻博客话题的突发性和时间性。

3.2 博客主题的开发

话题通常可以分为三种类型:稳定话题、短暂话题和波动话题。稳定话题可以持续一段时间,但很少有用户关注它。短暂话题往往会迅速消失在信息海洋中。波动话题很容易成为热门话题,本文重点关注具有波动特征的话题。根据以往的新闻研究,这样的话题可以经历自己的生命周期,包括诞生、成长、成熟和消亡阶段[50]。

一开始,一位博主赞助一期,然后吸引其他博主加入。当话题进入成长阶段时,由于早期新闻事件可能晦涩难懂、网络信息不可信,用户会主动发表意见或传播话题。在话题群体迅速壮大的同时,意见领袖

意见领袖是网络舆情的代表,具有舆情形成的影响力。如果话题存在争议,群体中往往会分裂成支持不同意见的派系。话题越有吸引力,意见领袖就越活跃。当话题发展到成熟期,舆情就会出现。如果大多数用户对事件的处理不满意,舆情就可能演变成其他观点。最后,意见领袖的影响力逐渐减弱,话题也就消失了。

根据上述描述,毫无疑问,博客空间中的用户、主题和观点是相互关联的。从W2T方法论的角度来看,人类相当于博客空间中的博客用户。网络相当于话题传播的上下文。信息粒度表示信息表达的语义层次。为了分析博客主题的形成和发展,从博客空间中提取了三个复杂网络。如图1所示,博客社区、话题网络和意见网络是话题传播的主要情境。博客社区由具有相似兴趣的用户组成。主题网络由不同粒度的主题以及主题之间的演化关系组成。意见网络由用户之间的意见和互动关系组成。

博客社区的边界可以通过博客话题来划分,而网络观点的演化与话题的演化息息相关,因此构建话题网络是分析话题传播过程中相关现象的关键。随着新闻事件的发展,话题网络的结构也随之发生变化。为了从微观和宏观两个角度表征博客话题的变化,信息粒度为

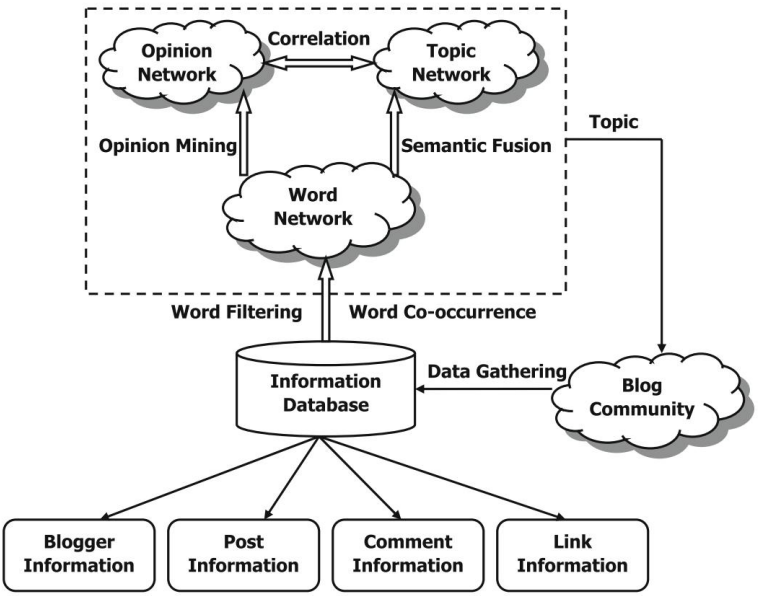


图1博客空间中的复杂网络

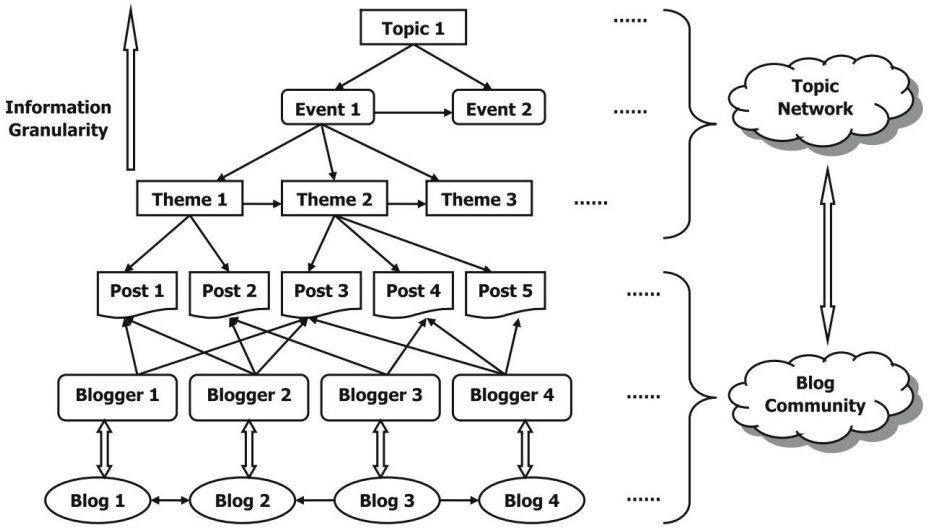


图2博客空间中主题网络的结构

考虑到这一点,主题网络的层次结构如图2所示。
不同层次的信息体现了人类信息处理的特点和话题传播的起点。

博客社区在不同的生命阶段具有不同的结构特征。根据社交网络分析理论,社区的结构与成员之间的互动模式和关系有关。图3和图4显示了2011年中国新浪博客网站博客热点话题的评论趋势。图3中的话题来源于中国国内事件,图4中的话题则相反。明显的现象是,大部分用户处于早期活跃状态。该现象可以解释如下。用户缺乏权威信息,对前期的事件非常好奇。

因此,在社区中就会出现羊群效应。当公众舆论

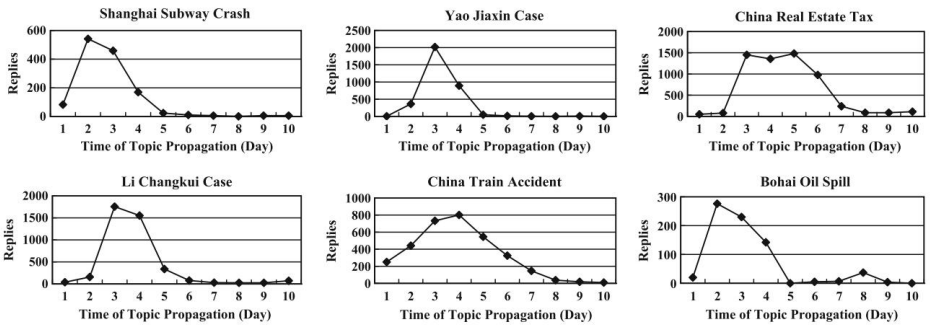


图3中国新闻事件相关热点话题评论趋势

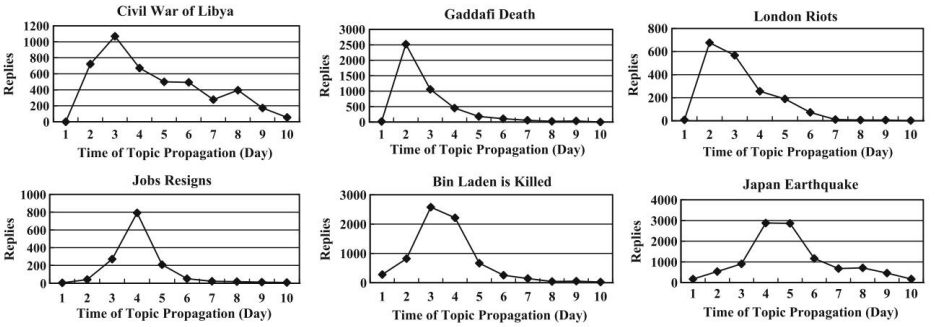


图4其他国家新闻事件相关热点话题评论趋势

一旦形成,沉默的螺旋就会出现。也就是说,当用户的意见处于少数时,他们往往会保持沉默。而且,随着时间的推移,话题的新颖性越来越弱,导致用户的行为模式发生巨大的变化。博客社区的结构变化能够反映话题的发展趋势,而用户的行为模式也决定了博客话题的突发性和时效性特征。就用户角色而言,用户分为三类,即意见领袖、活跃用户和普通用户。意见领袖在网络意见的演变过程中具有影响力。活跃用户扮演着传播者的角色。然而,与意见领袖相比,这种活跃用户的权威往往较弱。普通用户往往在舆论出现后潜伏。从职业背景来看,专家往往比普通用户发表的帖子多,很容易成为意见领袖。

此外,专业博主往往在工作日发帖,有的专业博主甚至把写作当成一份职业,而普通用户则多在周末发帖,目的只是为了消遣[53]。从以上分析可以看出,用户行为模式、网络舆情、意见领袖等均对博客话题的发展产生影响。

博客社区代表用户间的关系,观点网络代表这些关系的构建过程。对于一个热点话题,其网络观点往往在整个生命周期中经历较大的变化。因此,观点网络的结构变化可以反映用户的观点互动程度和网络观点的演化趋势。

此外,通过分析意见网络的特征,还可以识别意见领袖。例如,意见网络的结构和情感极性分布有助于在话题传播过程中识别意见领袖[8]。同时,意见网络的情感极性分布也反映了网络观点的发展趋势。

3.3 什么是意见领袖?

网络舆论演进过程中一个典型现象就是意见领袖的出现。意见领袖是博客用户,在影响其他用户的意见形成方面发挥着至关重要的作用。然而,博客空间中的意见领袖

与实体社区中的人不同。较高的知名度和认可度是意见领袖的基本条件。同时，普通用户并不太关注意见领袖在现实世界中的社会背景。此外，网络的传播速度加速了博客空间意见领袖的形成。博客空间中意见领袖的爆发性特征非常明显。根据以往的研究，意见领袖可分为波动型、长期型和短暂型意见领袖三种类型[44, 49]。

波动型意见领袖在网络舆情演进过程中经常变换自身角色，长期型意见领袖为了获得更好的认可度和热度，会积极参与话题互动；短暂型意见领袖在发表特定观点后，往往主动保持沉默或因周围反对而被动失去影响力。优质帖文数量是长期型意见领袖与短暂型意见领袖的主要区别。波动型意见领袖与长期型意见领袖相比，对提升认可度和热度兴趣不大，例如，一些公众人物在网络舆情出现问题时，经常参与舆情互动，以控制其他用户的情绪。

识别博客空间中的意见领袖是信息推荐和舆情监测的一项重要任务。对于用户角色的识别，通常会衡量用户对邻居或社区中其他用户的影响。

Song 等人通过测量帖子的重要性和新颖性来评估用户影响力[39]。Bodendorf等人通过社会网络分析发现意见领袖[8]。Akritidis 等人从用户行为的时间方面和相关帖子的质量来评估用户影响力[2]。Lim 等人首先根据用户的活动和使用模式构建社交网络，然后通过评估社交网络中帖子的质量来识别有影响力的用户[26]。

4 检测新闻博客中的热门话题

本节介绍了一种主题检测方法，该方法考虑了事件报告的不同视图以及事件之间的演化关系。在构建主题模型时还考虑了信息粒度。为了检测当前和即将到来的热门话题，除了用户兴趣之外，还考虑了每个话题的增长状态，因为增长状态不仅表明了话题的发展趋势，而且代表了话题的生命力。通过衡量话题的持续时间、话题新颖度、用户关注度以及话题增长度来评价话题的热度。

4.1 主题模型及热点话题检测算法

博客主题是一种Web信息，Web文档是重要的信息载体。在Web内容挖掘中，Web文档的信息提取、信息排列等信息处理需要仔细考虑信息粒度。对于Web文档的正文结构来说，节、段、句、词分别面向不同的信息粒度。因此，网络信息的清晰度与

信息粒度。在信息排列上,可以按照主题的类别将文档划分为不同尺度的簇,如果文档中的每一部分讨论的是某个具体问题或子主题,则可以将主题划分为问题或子主题。在用户表达上,由于个性化管理,用户可以从不同角度对同一主题发表评论,发表不同文风的帖子。因此,博客主题的结构可以具有多层次或多视角的特点。

考虑到语义表达中的信息粒度,主题可以被认为是由相关事件组成的簇,并且事件可以被认为是由相关主题组成的簇。对于主题的字面表达来说,可以用一组相关的关键词来表达主题。

因此,这个话题可以分为三层。三层模型如下:

– 主题模型定义为Topic = {Event1, Event2, ..., Eventn},其中Eventi表示第 i 个事件； – 事件模型定义为 Event = {Theme1,
Theme2, ..., Themem},其中
Themei表示第 i 个主题;
– 主题模型定义为主题= {关键词1,关键词2， ...,关键词},其中关键词i表示帖子的第i个关键词。

由于语义表达复杂,信息粒度很难衡量,有时不容易区分主题层和事件层的信息。因此,有必要分析主题和事件之间的区别。新闻博客主题通常源自事件或活动。对于自然灾害,事件报道的关键词涉及到灾害的早期情况。然后一些关键词转向灾害的处置和预防。与灾难相关的事件往往由于严格的人工控制而无法演变为其他事件,而主题模型实际上是面向整个事件的。例如,与事故相关的主题模型的构建如图5所示。箭头指向模型的下一个状态,这样的主题的发展体现在三个层次上。主题模型的演变反映了用户对地铁事故问题的转变,同一事件模型或主题模型也会发生相应的变化。然而,该活动是由总统竞选、奥运会等一系列相关活动组成的。

对于与奥运相关的话题,用户对不同赛事的关注程度不同

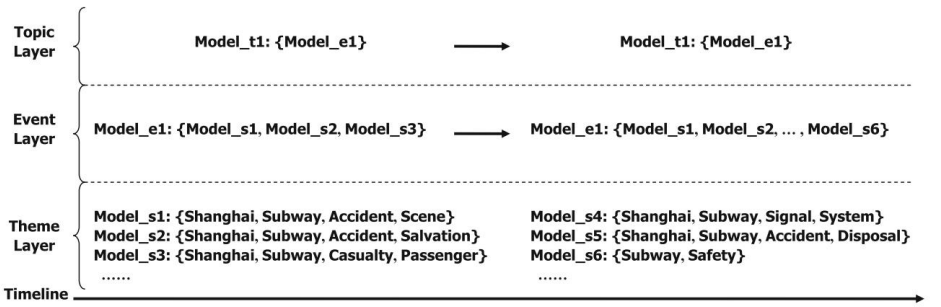


图5地铁事故相关主题模型构建

随着时间的推移,奥运会的备战情况、各类比赛情况、国家队在奥运会上的表现等报道纷纷出现。

而且活动面向的主题复杂,事件之间的演化关系需要仔细考量。

算法1中提出的热点话题检测算法可以分为三个部分,分别为话题检测（步骤1-13）、意见领袖识别（步骤14-17)和热度评估（算法1的其余部分）。由于新闻话题具有时间性,话题检测任务关注的是不同时间间隔内发布的帖子,因此采用帖子聚类的话题模型构建时间复杂度为O（n）。意见领袖的出现是指博主在一段时间内通过积累人气和认可度达到一定的影响力,因此意见领袖识别任务的最坏时间复杂度为O（n2）。话题热度评估任务关注的是特定时间间隔内活跃的话题,因此时间复杂度为O（n）。算法1由以上三个任务组成,时间复杂度为O（n2）。下面的三节将对这三个部分进行详细描述。

算法1热点话题检测输入： S为帖子集，
n为时间单位数， d为热点阈值。
输出： Tset是一个热门话题集。 1.对于S中的每
个帖子i执行2.从i中提取关键字； 3.构建i
的主题模型； 4.结束5.对于时间单位j = 0 到n执行
6。

根据相关关键词,找出与事件相关的帖子； 7.基于单次聚类策略构建事件模型； 8.将事件模型添加到j内的事件列表ELj中；
9. if j等于1 then手动构建j内的主题模型； else根据ELj中相关事件模型,构建新的主题模型或更新之前的主题模型； 13. end if 14.根据
用户互动关系,分析博客社区结构； 15.根据主题模型,检测每个主题上的网络观点； 16.构建博
客社区内的观点网络； 17.识别每个主题上
10. 的意见领袖； 18.统计回复者、意见领袖、回复和网络观点的变化
11.
12.

衡量j个主题内各个主题的成长状态；
19.通过测量持续时间、主题新颖性、注意力来评估j内的所有主题
用户和主题增长程度；
20.选择热度大于d的话题,添加新的热点话题到

Tset； 21.结束
22.返回Tset。

4.2 基于事件报告视角的话题检测

事件模型是构建主题模型的关键。新闻主题检测方法通常依赖于一些事件检测和跟踪的研究[33, 45]。列出了构建博客主题模型的两个重要任务。一是如何将传达不同粒度信息的相关帖子进行分组，以提取新闻事件。另一个是如何通过识别事件之间的演化关系来追踪新闻事件。单通道聚类算法广泛应用于事件检测和跟踪[3, 11]。基于事件报告视图的主题检测首先采用单通道聚类策略提取给定时间间隔内发生的事件。然后通过计算事件之间的内容相似度以及每个事件在不同主题上的分布来识别事件之间的演化关系。最后，通过检测新事件并跟踪先前事件来创建或更新主题模型。

基于事件报告视图的主题检测，针对按时间顺序排列的帖子，提炼出不同的主题模型。在主题检测过程中，采用不同的策略来处理不同层的信息。换句话说，主题、事件和主题模型是根据不同的策略构建的。以下三节将介绍如何构建这三种模型。 4.2.1节指出了如何从帖子中挑选关键词，以便以主题模型的形式表示帖子。单通道聚类算法对可能属于同一事件的帖子进行聚类。因此，后聚类意味着对主题模型进行比较，然后进行聚类以构建事件模型。 4.2.2节描述了主题模型之间比较的评估措施。事件之间演化关系的识别决定了主题模型的构建。 4.2.3节介绍了如何识别邻居事件之间的时间关系，从而创建或更新主题模型。

4.2.1 关键词提取和关键词关联

关键词是主题模型的基本元素。名词和动词从博客文章的标题和第一段中选择出来。文章的标签也被选为候选词。最后，根据候选词的权重确定关键词。在词频逆文档频率 (TFIDF) 方法中，候选词的权重通过以下公式计算[35]：

$$\text{权重}(tk,r) = T\,F(tk,r) * \log \frac{\text{否}}{NK + 0.5} \tag{1}$$

其中r是一篇博客文章， tk是一个单词， Weight(tk,r)是tk在r中的权重， TF (tk,r)是tk在r的文本中出现的频率， N是博客文章总数， Nk是出现tk的博客文章数量。

为了提高聚类算法的精度，采用信息检索策略提前过滤不相关的帖子。即，通过能够描述该事件的相关关键词来检索与特定事件相关的帖子。如果用户频繁发布和评论某个事件，则某些关键词在给定时间间隔内具有高度相关性。卡方检验

成功地用于衡量关键词之间的相关性,表示为
如下[6]：

$$\chi^2 = + \frac{(E(uv) - A(uv))2}{\text{紫外线辐射剂量}} \quad \frac{(E(uv^-) - A(uv^-))2}{E(uv^-)}$$
$$+\frac{(E(uv^-) - A(uv^-))2}{E(uv^-)} + \frac{(E(u'v') - A(u'v'))2}{E(u^-v^-)} \tag{2}$$

$$E(\text{紫外}) = \frac{A(u) \quad A(v)}{\text{氮}} \tag{3}$$

其中A(u)是单词u出现的帖子数， $A(u^-)$是单词 u 出现的帖子数
单词u未出现的帖子中， A(uv)是包含单词 u 的帖子数量
u和v都出现在其中， N是帖子总数。如果一个关键词集合
包含至少两个关键词,每个关键词对都是随机提取的
恢复。

4.2.2 后聚类

单次通过聚类算法可以解释如下。报告合并
如果两份报告之间的内容相似度高于阈值,则认定为事件。
否则,该报告将被视为新事件的第一个报告。什么时候
文本以单词集表示,杰卡德系数计算
两个集合之间的重叠率通常用于评估
两个文本[19, 42]。杰卡德系数越大，
两段文本之间的相似度为。然而,如图6所示,Jaccard系数
没有考虑集合的大小,文本聚类效果不好
两组不处于同一级别。在我们的方法中,帖子用其自己的
主题模型,基于相似度来评价帖子之间的相似度
主题模型之间。因此,所提出的方法比较了差异
根据 Jaccard 系数计算两组之间的大小,以及以下内容
方程用于比较两个帖子之间的相似度：

$$Sim(di, dj) = \alpha * + \beta * \frac{|di \cap dj|}{|di \cup dj|} \quad \frac{|di \cap dj|}{\text{分钟}(|di|,|dj|)} \tag{4}$$

其中， dj表示第 j 个主题的关键词集合， Sim(di, dj)为相似度
第 i 个主题和第 j 个主题之间， di ∩ dj是集合di和dj 的交集，

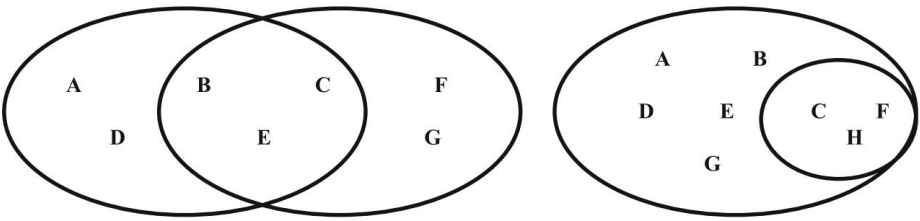


图6两个集合的交集

di ∪ dj是集合di和dj 的并集， |di|是集合 di 的大小， α 和 β 是系数， min(y, z) 是y和z之间的最小值。

4.2.3 事件间演化关系的识别

图7给出了事件演化的两种情况,左图表示一个事件可能演化为另外两个事件,右图表示多个事件可能演化为一个新事件。事件的属性有时会发生很大变化,因此需要仔细考虑事件之间的演化关系。对于Hong等人的工作,将主题和报告划分为一些子主题,然后根据相关子主题的比例和分布来检测新主题[18]。如果相关子主题的比例较低,且此类子主题的分布分散,则将报告归类为新主题。另一方面,在评估事件之间的关系时,除了事件之间的内容相似性之外,事件发生的时间也是另一个关键因素。如果两个事件发生的时间非常接近,则事件之间很可能存在相关性。相反,即使它们之间的内容相似度很高,事件也不太可能相关。本文提出的方法采用了Hong的策略。根据事件间内容相似度,找出相邻时间间隔内与目标事件相关的主题,再根据相关主题的分布情况,识别事件间的演化关系。 识别两个事件间的演化关系的具体过程如下：

1、识别原始事件Eventi所属的主题模型Topick； 2、根据事件之间的内容相似度,针对目标事件Eventj(j = i)构建前一个时间单位内的相似事件集合Sij(i = j)； 3. 如果Sij不为空,且Sij中的大部分事件属于Topick，则存在

Eventi和Eventj之间的进化关系； 4、若Sij为空,则构建当前时间单位内发生与Eventj相似的事件且早于Eventj的相似事件集合S； 5. S ij的策略与上述相同

伊

被采纳。

如果目标事件只有一个相似事件,且内容相似度非常高,则认为两个事件是同一事件。

根据事件模型的定义,主题模型是事件模型的基础。因此,比较事件之间的内容相似度意味着需要比较相关的主题模型。当两个事件进行比较时,每个事件

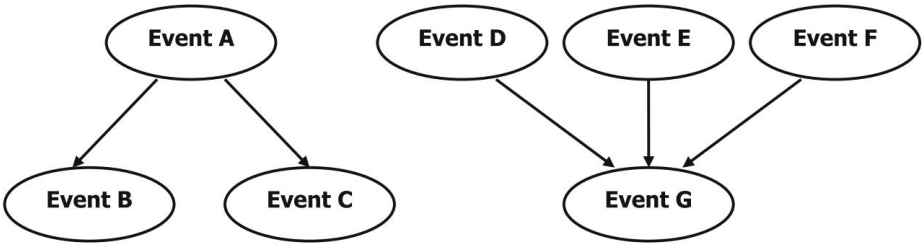


图7不同案例中事件的演化关系

将一个事件模型中的主题与另一事件模型中的主题进行比较,并使用以下事件相似度方程来衡量两个事件之间的内容相似度：

$$Comp(ei, ej) = \frac{1}{厘米} \sum_{p=1}^{中文} \sum_{q=1}^{厘米} Sim(dip, djq) \quad (5)$$

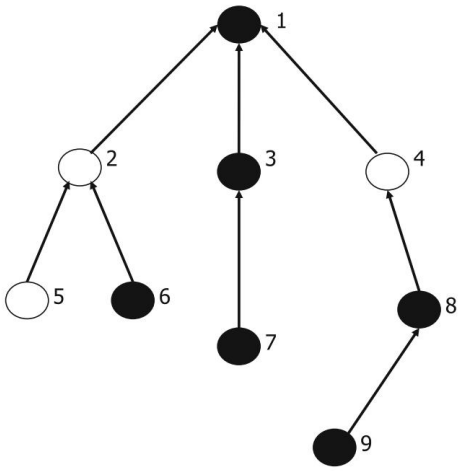
其中， Comp(ei, e j)表示事件ei与e j的相似度， cn表示属于事件 ei 的主题数量， cm表示属于事件e j的主题数量， dip表示事件 ei 的第 p 个主题， djq表示事件e j的第 q 个主题。

4.3 意见领袖识别

意见领袖的识别是判断博客话题发展趋势的重要指标,意见领袖具有较高的知名度和认可度,从社会网络结构上看,意见领袖相对于其他用户具有较大的中心度,因为意见领袖拥有一批追随者,而且他们对社会现象有深刻的洞察,使他们的观点具有代表性。

用户对网络意见的影响可以通过统计邻居意见的情感极性分布来衡量。因此,意见网络对于识别意见领袖很有用。基于用户评论或帖子引用的博客社区是构建这种意见网络的重要前提。此外,基于主题模型提取用户意见,并分析每个意见的情感。在此基础上,构建了不同时间区间的意见网络。如图8所示,节点代表用户的意见,节点之间的边代表用户之间的交互关系。此外,黑色节点代表正面意见,白色节点代表负面意见。每条边的箭头都指向被注释的对象。

图8基于观点交互关系的观点网络



4.3.1 意见提取和情感分析

意见挖掘对于意见领袖识别非常重要,其组成包括
主题提取、意见持有者识别、主张选择和情感
分析。主题提取重点在于识别评论对象。
然而,注释对象面向不同的事件属性,并且
有时一个用户可能会批评另一个用户。因此,主题提取需要
考虑观点互动的语境,避免话题
漂移。而且,用户回复是任意的。例如,有一个完整的句子,
没有主语和情感词的句子。因此,对互动的认识
当用户回复是
没有主语或情感词的句子。

一般来说,名词、动词、形容词和副词对情感有用
网络舆论分析。句子中的停用词因此被过滤
以保证情感分析的准确性。 Ku等人提出的方程
等人。 [22]用于识别情感极性的描述如下：

$$S(T) = \sum_{i=1}^{\text{否}_{\text{数数}}} \text{科学} \tag{6}$$

$$Sci = Pci - Nci \tag{7}$$

$$PCI = \frac{\sum_{j=1}^{n} fpcj}{\sum_{j=1}^{n} \text{韓式}_{\text{膠囊機}} + \sum_{j=1}^{*} f\,nci\,f\,ncj} \tag{8}$$

$$\text{国家癌症研究所} = \frac{\sum_{j=1}^{*} NCJ\ NCJ}{\sum_{j=1}^{n} fpci + \sum_{j=1}^{*} f\,nci\,f\,ncj} \tag{9}$$

其中S(T)表示句子 T 的情感极性， Ncount是
在句子T中不属于停用词的词中， Sci是情绪
句子T中第i个单词的极性， fpci表示第i个单词的频率
句子T在正面意见样本集中的频率， f nci表示该句子出现的频率
负面意见样本集中句子T的第i个词， n是总的
正面意见样本集中的不同词语， m为不同词语的总数
负面意见样本集中的单词， j表示意见中的第 j 个单词
样本集， fpcj为正面意见样本集中第j个词的出现频率，
f ncj为负面意见样本集中第j个词的出现频率。

4.3.2 帖子排名

帖子质量往往是根据帖子之间的链接关系来衡量的。
然而,链接有多种类型,博主为不同的目的设置不同的链接
社交或信息推荐等不同目的。因此，
主题漂移通常会对帖子排名算法产生负面影响,如果链接
信息不被分析。事实证明,帖子的质量很可能是

如果帖子有长文本、大量评论或内链接，而外链接则少得多[1]。因此，本文分析了链接的类别，并为不同类型的链接分配了不同的权重。将帖子链接到博客的链接与帖子之间的链接不同。帖子之间的超链接根据内容进行过滤。基于Agarwal方法的后影响方程描述如下：

$$I(pa) = wlen * Len(pa) * wcom * Rp(pa) + wqu * Tr(pa) + 获胜 * \sum_{我=1}^{*} I(pi) - wout * \sum_{j=1}^{n} 我(pj) \quad (10)$$

其中， I(pa)为帖子pa的影响力， Len(pa)为帖子pa的长度， Rp(pa)为帖子pa的回复数， Tr(pa)为帖子pa的引用数， wlen为帖子长度系数， wcom为评论系数， wqu为引用系数， win为入链系数， wout为出链系数， i表示链接到pa的第i个帖子， j表示pa链接到的第j个帖子， m为帖子pa的入链总数， n为帖子pa的出链总数。

4.3.3 用户影响力评估

李等人的研究表明，通过评估用户影响力可以发现有影响力的博主，而影响力分为社交量、发帖内容和活跃度三个方面[25]。本文提出的方法基于李的评估策略，通过评估博主对网络舆情的影响力，从有影响力的博主中识别出意见领袖。引用是网络社区用户影响力的主要体现。因此，本文构建了基于帖子间引用的信息传播网络，并计算代表博主的节点在网络社区中的中心度。博主的引用次数越多，社交量的用户影响力就越大。同时，意见领袖往往通过相关帖子引领共识的方向。然而，由于信息过载，浏览者无法阅读完整的帖子。因此，博主必须尽力发布更多高质量帖子，以影响更多的用户。帖子内容的用户影响力通过用户高质量帖子占总帖子数的比例和主题组中支持者的比例来评估。对于用户影响力的活跃度，可以通过参与话题讨论来提升影响力。

因此，通过统计用户在观点互动过程中发表的、代表用户观点的回复数量来评估用户活跃度的影响力。

识别意见领袖的影响力评价方程定义如下：

$$意见n(b,x) = \sum_{我=1}^{n} \frac{nopi(b,x)}{特诺皮(x)} \frac{nosi(b,x)}{} b\ pi(b,x) \quad \psi \quad \frac{ceni(b) + \quad + \delta}{氮氧化物(x)} \quad \frac{}{时间点\ (x)} \quad (11)$$

其中Opindn(b, x) 为用户b在第 n 个时间单位内对主题x的影响力， i表示第 i 个时间单位， nopi(b, x) 为用户 b 在第 n 个时间单位内对主题 x 的意见数

用户b在第i个时间单位内针对主题x发表的帖子数， Tnopi(x)为所有用户在第i个时间单位内针对主题x发表的帖子总数， ceni(b)为用户b在第i个时间单位内在社会网络中的中心度， nosi(b, x)为第i个时间单位内与用户b观点相同的用户数量， Tnosi(x)为第i个时间单位内针对主题x发表帖子总数， bpi (b, x)为用户b在第i个时间单位内针对主题x发表的高质量帖子数， Tpi(x)为所有用户在第i个时间单位内针对主题x发表的帖子总数,ψ为区位系数,为情感系数,δ为引用系数。

4.4 话题热度评估

新闻博客的特点表明,用户兴趣起着重要作用。而且,主题的生命力是动态的,而时间主题的生命周期尤其短。因此,通过测量用户的兴趣程度和话题的成长程度来检测热点话题。

就用户兴趣而言,话题传播的情境可以刺激用户参与话题互动。如果一个话题传播的时间较长,那么它就有很大概率吸引用户。如果一个话题很新,博客用户也更有可能对该话题感兴趣。如果与某个话题相关的帖子被引用次数高或者回复次数多,那么该话题很容易被网站推荐给所有用户。因此,如果一个话题满足上述所有条件,那么用户对该话题的兴趣无疑较高。因此,用户对某个话题的兴趣可以通过衡量话题的时长、话题的新颖性和用户的关注度来评估。本文提出的方法从人类的遗忘因素的角度来评估话题的新颖度[4]。即,如果事件发生的时间较近或者发生的频率较高,那么人类可以清楚地记住该事件。反之,如果事件发生的时间较长或者很少发生,那么人类的记忆就会衰退。通过统计某个话题拥有的帖子被引用和回复的总数来评估用户的关注度。

为了检测话题活力的变化,需要定期对话题增长情况进行评估。根据3.2节对话题发展的分析,网络共识的出现意味着博客话题进入成熟期,这是一个重要的转折点。相关的传播学研究发现,满足以下条件即可形成网络共识[28]。第一,回复数达到一定水平。

二是意见领袖积极发挥作用。三是用户情绪被点燃。
第四,大量媒体参与话题传播。张文等的工作忽略了网络共识对网络话题的影响,通过点击量和回复量来衡量话题增长[51]。本文虽然采用了张文等的衡量策略,但所提方法根据不同时间间隔内回复量、回复者、意见领袖和网络观点的变化来衡量话题增长。统计观点时,忽略同一用户发表的重复观点。话题增长可以用以下公式来衡量：

$$\text{增长}\ (x) = \mu 1 \sum_{\text{我}=1}^{n} f(mi) + \mu 2 * \sum_{\text{我}=1}^{n} f(li) + \mu 3 * \sum_{\text{我}=1}^{n} f(ci) + \mu 4 * \sum_{\text{我}=1}^{n} f(si) * \frac{1}{n} \quad (12)$$

$$f(pi) = \begin{cases} 1, & 我= 1 \\ 范数（），\overline{pi-1+0.1} & 我> 1 \end{cases} \quad (13)$$

$$范数(y) = \begin{cases} 1， & y⩾1 \\ y,否则 \end{cases} \quad (14)$$

$$\mu1 + \mu2 + \mu3 + \mu4 = 1 \quad (15)$$

其中Growth(x)是主题x的增长程度， mi是回复者数
主题x在第i个时间单位内， li为主题x的意见领袖数量
第i个时间单位内， ci为第i个时间单位内主题x的回复数，
si为第 i 个时间单位内对主题x的评论数， n为总数
时间单位， μ1是用户的增长系数， μ2是
意见领袖， μ3为回复增长系数， μ4为增长系数
的一个意见。主题的成长度从0.0到1.0不等,主题处于
如果该值接近 1,则为成熟阶段。

主题热度通过以下公式进行评估：

$$热度(x) = \frac{铜}{n} *增长(x) * (\lambda * sc(x) + xi * qu(x)) * t(x)^{-k} \quad (16)$$

其中Hotness(x)表示主题x的热度， n是时间单位的总数，
cu为主题x出现的连续时间单位数， Growth(x)为
主题x的增长程度， sc(x)为主题x的回复总数， qu(x)为
属于主题x的帖子的引用次数， t(x)是时间差
主题x的发布日期和当前时间之间,λ是评论
系数,ξ为帖子的引用系数， k为衰减系数。

5 实验与讨论

为了验证其可行性和有效性,进行了实验
所提出的方法。同时,意见领袖对话题的影响力
传播及其特性进行了分析。测试样本集包括1520
在中国新浪博客网站上发表了 202290 条相关帖子和 202290 条回复[37]。
测试样本的发布日期为2011年11月9日至2011年1月18日之间，
2012. 训练样本集包括17910个纯文本,用于关键词提取
来自中国搜狗实验室[38]， 14317 条网络评论
分析,以及中国新浪博客网站列出的12个博客热点话题
2011 年,涉及社会、政治和经济。ICTCLAS 软件是
应用于中文分词[20]。

5.1 所提出方法的性能

命名实体识别是自然语言处理中的一个大问题。
博客数据的整理也存在同样的问题。为了提高ICTCLAS命名实体识别的精度,构建了用户词典

手动根据帖子的标签,因为这样的标签可以成功使用

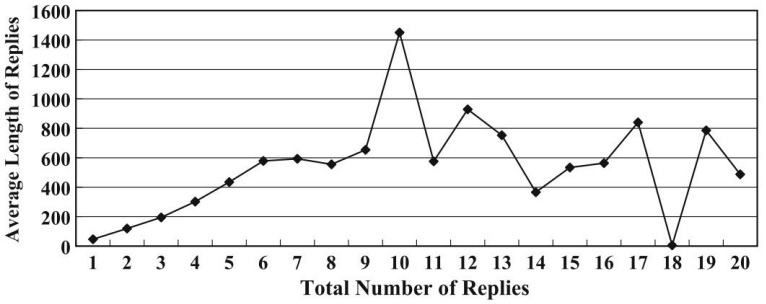


图9假名用户回复特征

检测突发事件[46]。此外，匿名用户频繁参与意见互动，更有可能发表负面意见[21]。

因此，了解博客社区的结构需要分析匿名用户的特征。事实证明，大多数博客作者倾向于使用用户名而不透露真实姓名[21]。因此，可以通过观察使用假名的用户的行为模式来推断匿名用户的特征。按回复总数统计不同类型的假名用户的回复次数和平均回复长度。如图9所示，训练样本集中的假名用户在之前的回复较长时，更有可能继续对该主题进行评论。因此，设置一个评论阈值来估计主题组内匿名用户的范围，其值设为200。

实验一在所提方法的基础上评估话题热度，但不考虑话题的增长程度。实验二采用所提方法评估话题热度。实验三采用基于凝聚层次聚类算法的热点话题检测方法[14]，通过统计发帖总数和回复总数来评估话题热度。为了选取最优参数来提高所提方法的性能，首先观察参数调节对性能的影响。如图10所示，评论系数λ设置为0.1，帖子引用系数ξ设置为0.5。如图11所示，所提方法的性能

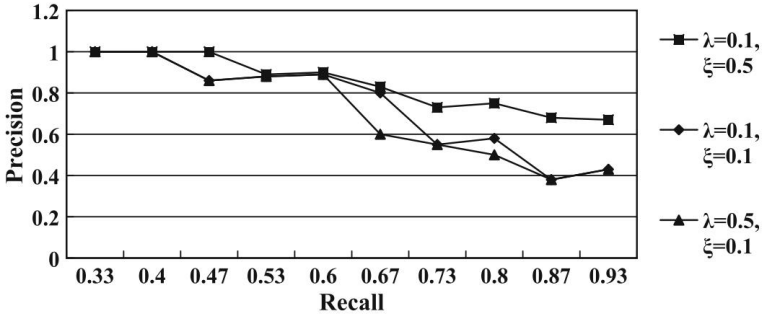


图10所提方法不同参数下的性能比较

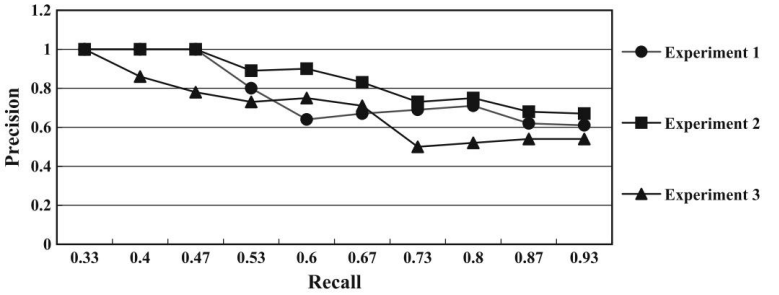


图11三个实验的性能对比

凝聚层次聚类算法基于向量空间模型,而大部分博客文章并未经过归一化处理,导致基于凝聚层次聚类算法的话题检测方法准确率较低。而基于事件报道视角的话题检测侧重于文章关键词,因此未归一化的文章对本文方法影响不大。另外,通过衡量博客话题的增长状态可以提高热门话题检测的准确率。另一方面,博客网站推荐的话题在三次实验中均被检测到,因此博客服务商对话题传播的影响不容忽视。

回复者、回复者、意见领袖和网络意见是评价一个博客话题发展趋势的四个重要因素。传统方法常常忽略意见领袖和网络意见,意见领袖和网络意见对所提出方法性能的影响需要验证。因此,考虑以下三个条件。

条件1在使用4.4节中列出的主题热度评估方程评估主题时考虑了上述四个因素。条件2忽略了意见领袖的影响。条件3忽略了意见领袖和网络舆论的影响。性能对比如图12所示,综合考虑这四个因素后,热度评估结果良好。

争议中经常出现话题漂移,很多回复对于评价话题热度并无帮助。相反,网络对事件属性或事件的看法可以准确地反映一个话题的状态。此外,意见

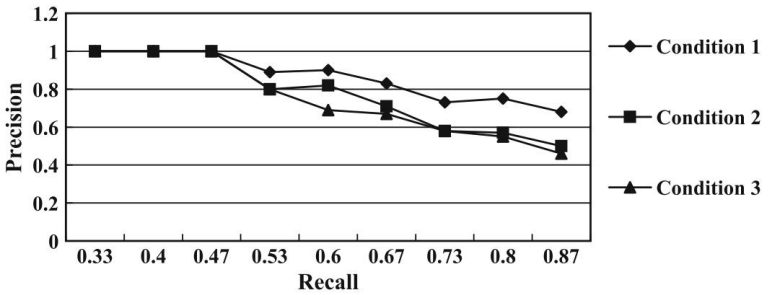


图12三种条件下的性能比较

意见领袖在网络舆论形成过程中起到至关重要的作用,其意见具有代表性,如果将意见领袖也考虑进去,效果会更好。

5.2 关于意见领袖的讨论

意见领袖必须积极发表高质量的帖子,才能维持地位。
因此,衡量帖子的质量对于认可意见领袖非常重要。为了测试4.3.2节所述改进的帖子排名方法,从训练样本集中提取了与9个热门话题相关的378个帖子,其中包括新浪网站推荐的86个帖子。

Agarwal 等人提出的方法。 [1]用作基线。实验结果如图13所示。虽然有些博主关闭了帖子回复功能,但改进方法考虑了帖子引用来抵消缺乏评论功能的影响。此外,帖子之间的链接经过区分和过滤,以避免主题漂移。因此,改进的后排序方法具有更好的性能。

某一领域的知名博主往往拥有大量的关注者,比普通用户更有优势。为了识别知名博主是否是意见领袖,利用新浪博客网站选取的14个领域的1406名知名博主进行实验,将出现在训练样本集中的意见领袖识别为:使用第 4.3.3 节中所述的影响评估方程。观察意见领袖和著名博主在不同时间间隔的行为。如图14所示,不同时间间隔的社交网络是根据意见领袖和著名博主之间的互动关系构建的。社交网络中的圈子节点代表当前时间区间内的意见领袖。矩形节点代表著名博主或以前的意见领袖,箭头指向响应者。

图14代表了本拉登被击毙新闻相关话题发展过程中意见领袖和知名博主的行为模式。

这些实验结果表明,与其他类型的意见领袖相比,长期意见领袖属于少数。特定领域的著名博主并不总是意见领袖。意见领袖更喜欢通过博文的方式影响其他用户,因为他们很少参与其他博客的意见互动。分析意见领袖在舆论引导过程中的影响力。

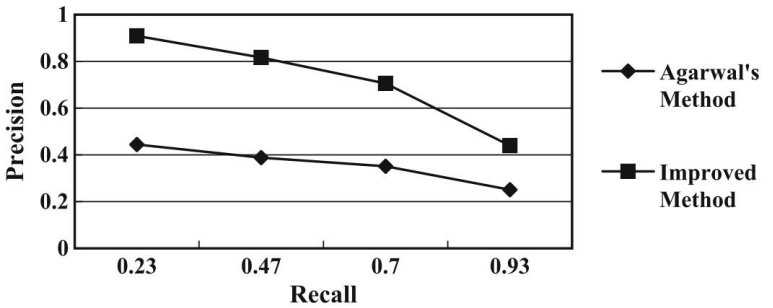


图 13两种方法在帖子排序上的表现比较

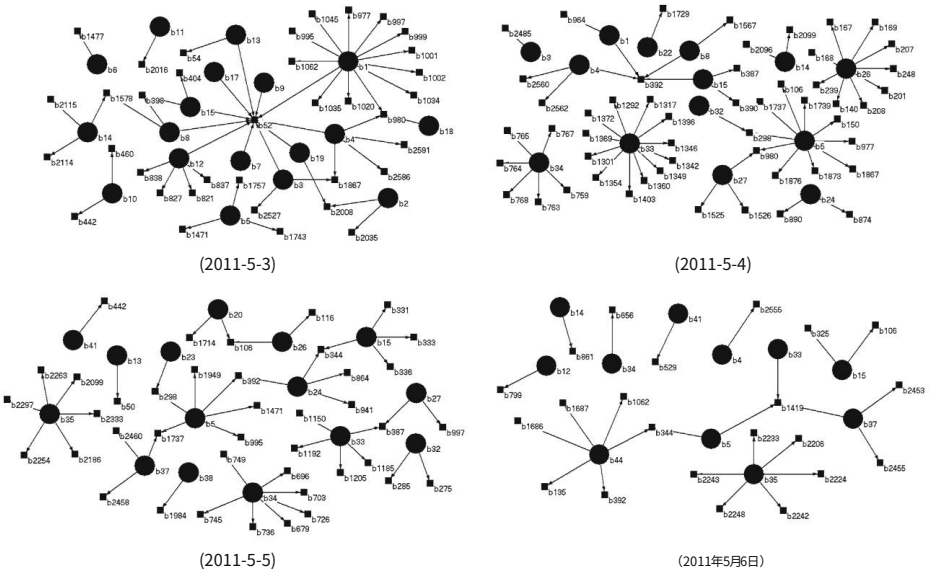


图14不同时间段内意见领袖与知名博主的互动情况

主题传播，构建不同时间间隔内的主题传播网络。在当前或前一阶段发布原始报告并扮演意见领袖角色的信息传播者以圆形节点的形式进行标记，矩形节点代表网络中的信息接收者。

博文的引用量用一条连线表示为传播者和接收者之间的连线。图15展示了本拉登被击毙这一新闻话题的传播过程，各个传播者的帖子引用量呈现下降趋势。实验结果显示，意见领袖在话题传播过程中的影响力逐渐减弱，因此意见领袖必须及时掌握最新资讯，发表犀利言论，才能保持自己的地位。

上述现象的原因可以解释如下。博主的真实身份被忽略，信息的可信度基于合理的解释和相关事件的发展状况。由此可见，意见领袖的形成呈现出快速且不稳定的特点。同时，随着检索工程师的提高和大量网络媒体的关注，用户很容易了解全面的信息。

此外，有关部门往往会主动公布热点事件的最新调查结果，意见领袖对话题传播的影响力相对较弱。

意见领袖的地位是在信息传播过程中形成的。对于意见领袖来说，如何让其他用户接收并接受他/她的信息是一个重要的问题。因此，影响用户信息选择的因素变得越来越重要。至于Consuantiou和同事的工作，采用不同的启发式方法从推荐和读者的百分比、信息来源、文本和图片的特征来了解用户对在线新闻的选择[13]。他们的研究发现，百分比

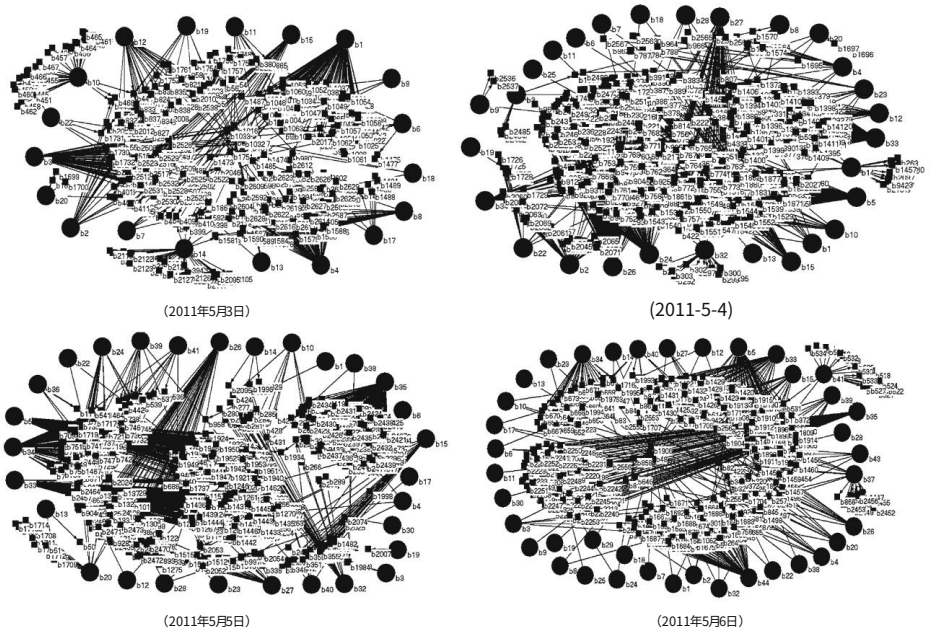


图15本拉登被击毙事件相关热点话题传播过程

推荐和读者的数量以及信息资源的声誉对用户的选择有积极的影响。然而,我们的实验结果有一些不同的发现。我们的研究通过基于帖子引用的信息传播网络来观察意见领袖对话题传播的影响。我们的实验结果表明,博主的引用行为主要发生在非常有限的时间内,并且随着时间的推移,帖子的影响力逐渐变弱。另一方面,博主很少同时引用许多相关的帖子,大多数博主更愿意选择拥有相同兴趣和观点的博客。因此,我们的研究表明,主题的新颖性对用户的选择有很大的影响,而拥有高推荐率和读者群的帖子通常不会长期影响博主。用户的兴趣也对用户的选择有很大的影响,这可以解释为什么博主有时更喜欢新闻博客而不是权威来源。

6 结论和未来工作

信息过载是Web挖掘的一个大问题。然而,热点话题的一些代表性特征可以在一些相关博客中体现出来,而博客话题的传播机制使得用户在话题的形成和发展中发挥着至关重要的作用。因此,基于W2T方法来分析这些相关博客中出现的现象来检测热点话题是很有用的。新闻博客的主题与新闻事件和用户的个人愿望密切相关。换句话说,这样一个话题的内容是由一个事件的叙述和

博主的博文主题体现了博主个人的意愿或观点,博主往往从不同的角度或层次发表对同一新闻事件的主题观点,因此博文信息被处理在不同的信息粒度上。时间性是新闻博文主题的典型特征,在不同的生命周期阶段对主题热度的评价结果不同,因此判断博文主题发展趋势的因素十分重要。意见领袖的影响力被网络平台放大,甚至影响相关新事件的发展,因此,了解意见领袖的形成特征有助于正确分析其形成机制,有效评估其影响力。本文的主要贡献如下：

– 提出了一种基于事件报告视图的主题检测方法,考虑事件之间的信息粒度和演化关系来构建新闻博客主题模型。

– 提出了一种博客空间中意见领袖的识别方法,该方法可以精细地衡量博客用户的受欢迎程度和认可度。

– 提出话题增长状态评价指标,用于检测当前和未来的热点话题,考虑回复者、意见领袖、回复和网络观点的变化。

实验结果证明了所提方法的可行性和有效性。
但在今后的工作中还存在一些不足。首先,实验不考虑同义词。其次,事件之间的演化关系可以进一步分解为时间关系和因果关系,而不考虑事件之间的因果关系。第三,有些帖子很受欢迎,因为帖子中嵌入了一些令人印象深刻的图片或视频。

因此,需要分析图片或视频对用户的影响。最后,尽管虚拟网络打破了物理世界的一些障碍,但在物理世界中形成的人性仍然对虚拟网络产生很大的影响。因此,未来需要进一步研究博客空间中不同用户的行为模式,可以采用社会学或心理学的理论来理解用户行为。同时,还需要很好地分析用户的个性,因为用户的行为反映了他/她自己的需求和兴趣。而且,信息在传播过程中会被用户重新加工、提炼。如果按照人类信息处理机制来提取这些信息,解决问题的策略可能会更加有效和高效。未来文字处理的不足将会得到改善。此外,我们将更加关注博客空间社交联系模式的最新调查。

致谢本研究得到国家自然科学基金项目（60905027)和北京市自然科学基金项目（4102007)的资助。

参考

1. Agarwal, N.、Liu, H.、Tang, L.：识别社区中有影响力的博主。见：网络搜索和网络数据挖掘国际会议论文集,第 207-217 页（2008 年）

2. Akritidis, L.、Katsaros, D.、Bozanis, P.:识别博客网络中富有成效和影响力的博主社区。 IEEE 传输。系统。曼赛博恩。 41(5), 759–764 (2011)
3. Allan, J.、Papka, R.、Lavrenko, V.:在线新事件检测和跟踪。载于:第二十一届国际 ACM SIGIR 会议论文集,第 37-45 页（1998 年）
4. Anderson, JR, Schooler, LJ: 记忆中的环境反射。《心理学》 2(6)，396–408 (1991)
5. Balakrishnan, H., Deo, N.:在复杂网络中发现社区。在:会议记录第四十四届东南地区年度会议,第 280–285 页（2006 年）
6. Bansal, N.、Chiang, F.、Koudas, N. 和 Wm, F.:在博客圈寻找稳定的集群。载于:第三十三届超大型数据库国际会议论文集,第 806–817 页（2007 年）

7. Blei, DM、Ng, AY、Jordan, MI:潜在狄利克雷分配。J. Mach. Learn. Res. 3,993–1022 (2003 年)
8. Bodendorf, F.、Kaiser, C.:检测在线社交网络中的意见领袖和趋势。第四届数字社会国际会议论文集,第 124-129 页（2010 年）
9. Brants, T.、Chen, F.、Ioannis, T.:基于主题的文档分割与概率潜在语义分析。载于:第十一届国际信息和知识管理会议论文集,第 211-218 页（2002 年）

10.曹YZ,邵PJ,李LQ。博客网络中基于扩散阈值的话题传播模型。见:2011 年商业计算和全球信息国际会议论文集,第 539–542 页 (2011)

11. Chen, CC, Chen, YT, Chen, MC:事件生命周期建模的老化理论。 IEEE 传输。系统人机控制论37(2), 237–248 (2007)
12. Chen, KY, Luesukprasert, L., Chou, SCT:基于时间线分析和多维句子建模的热点话题提取。 IEEE 传输。知道。数据工程19(8), 1016–1025 (2007)
13. Constantiou, L.、Hoebel, N.、Zicari, RV:框架策略如何影响用户对网络内容的选择。并发计算。练习。经验。 24(17), 2207–2220 (2012)
14. 戴XY,陈QC,王XL,徐J.:基于层次聚类的财经新闻在线主题检测与跟踪。见:第九届机器学习和控制论国际会议论文集，卷。 6,第 3341–3346 页（2010 年）

15. 丁锋: 网络社区信息互动与传播研究. 北京北京交通大学（2010）
16. 龚慧晶:网络热点话题自动检测研究。华中师范武汉大学（2008）
17. He, TT, Qu, GZ, Li, SW, Tu, XH,zhong, Y., Ren, H.:半自动热点事件检测。载于:第二届高级数据挖掘和应用国际会议论文集,第 1008–1016 页（2006 年）

18. 洪英,张燕,范建玲,刘婷,李胜:基于划分比较的新事件检测子主题。计算机学报31(4), 687–695 (2008)
19. Huang, HH, Kuo, YH:跨语言文档表示和语义相似性测量基于模糊集和粗糙集的方法。 IEEE 传输。模糊系统18(6), 1098–1111 (2010)
20. ICTCLAS。主页： http://ictclas.org。访问日期:2011 年 3 月 10 日 21. Kilner, PG, Hoadley, CM: 在线实践社区中的匿名选择和专业参与。在:2005 年计算机支持协作学习会议论文集,第 272-280 页 (2005)

22. Ku,LW,Liang,YT,Chen,HH:新闻和博客语料库中的意见提取、总结和跟踪。载于:AAAI-2006 网络日志计算方法春季研讨会论文集,第 100–107 页 (2006)

23. Kumar, R.、Novak, J. 和 Raghavan, P.: 论博客空间的爆发式发展。万维网8(2), 159–178 (2005)

24. 李俊杰,张新昌,翁勇,胡希杰:基于文本观点分析的博客热度评价模型。载于:第八届 IEEE 国际可靠、自主和安全计算会议论文集,第 235–240 页（2009 年）

25. Li, YM, Lai, CY, Chen, CW:发现博客圈营销的影响者。信息。Sci. 181(23), 5143–5157 (2011)
26. Lim, SH, Kim, SW, Park, SJ, Lee, JH: 确定博客网络中的内容超级用户:一种方法及其应用。IEEE 系统人机控制论汇刊 41(5), 853–862 (2011)
27. Liu, Y.、Yu, XH、An, AJ、Huang, XJ:顺应情绪变化的潮流:利用不断发展的在线评论进行情绪分析。万维网。doi :10.1007/s11280-012-0177-1

28. 罗华: 社会焦点事件网络舆情演化研究.华中科技大学学报, 2015, 34(5): 774-778.
武汉科技大学（2011）
29. 马新华、李丽:人们为什么写博客?探索写博客的动机。在:诉讼程序
第二届 IEEE 网络社会研讨会,第 119–122 页 (2010)
30. Musial, K.、Budka, M.、Juszczyszyn, K.:在线社交网络的创建和发展如何
社交网络进化?全球资讯网。 doi:10.1007/s11280-012-0179-z
31. Musial, K., Kazienko, P.: 互联网上的社交网络。万维网16(1), 31–72 (2013)
32. Pan, X.:复杂网络上的意见传播模型。大连理工大学
大连（2010）
33. 齐红锋: 网络舆情热点话题检测与事件追踪研究.哈尔滨
哈尔滨工程大学 (2008)
34. Qiu, HM:博客圈的社交网络分析。哈尔滨工业大学,哈尔滨
（2007 年）
35. Salton, G., Buckley, C.:自动文本检索中的术语加权方法。信息
加工与管理24(5), 513–523 (1988)
36. 石建军,胡梅,戴光哲:基于小世界模型的中文文本主题分析。
中文信息处理21(3), 69–75 (2007)
37. 新浪博客网站。主页： http://blog.sina.com.cn。 2012 年 2 月 1 日访问 38. 搜狗实验室。主
页： http://www.sogou.com/labs/dl/c.html。 10 月 28 日访问
2009年
39. Song, XD、Chi, Y.、Hino, K. 和 Tseng, B.:识别博客圈中的意见领袖。
出处:第十六届 ACM 信息与知识管理会议论文集,第 971-974 页（2007 年）

40. Sun,WJ,Qiu,HM:博客圈的社交网络分析。载于:2008 年管理科学与工程国际会议论文集,第 1769–1773 页 (2008)

41. 王春华、张梅、马胜平、茹丽燕:网络环境下的在线新闻自动生成，《第十七届万维网国际会议论文集》，第 457-466 页
（2008 年）

42. Wang, JH: 基于 Web 的单篇短文档词条代表性验证. 2011 年 IEEE/WIC/ACM Web 智能与智能代理技术国际会议论
文集, 第 3 卷, 第 114–117 页 (2011)

43. 王颖,席永华,王玲:基于改进的K均值分割挖掘中文网页热点话题,第八届机器学习与控制论国际会议论文集,第255-260
页（2009年）

44.谢光华:网络意见领袖影响力系统研究.中原
武汉师范大学（2011）
45. Yang, CC, Shi, XD, Wei, CH: 从新闻语料库中发现事件演化图。IEEE 系统人机控制论汇刊。39 (4), 850–863 (2009)

46. Yao, JJ, Cui, B., Huang, YX:协作标签的突发事件检测。万维网15(2), 171–195 (2012)

47. Yao, JT, Yao, YY: 基于 Web 的信息检索支持系统的信息粒度。《光学仪器工程师学会会刊》第 5098 卷,第 138-146 页
（2003 年）

48. Yao, YY、Petty, S.:网络内容的多种表示形式以实现有效的知识利用。
见:2012 年国际脑信息学会议论文集,第 338–347 页（2012 年）
49.余华:政治论坛意见领袖研究　以中日强国论坛论坛为例。华中科技大学,武汉（2007）

50. 张燕:BBS舆情传播现象研究.吉林大学,长春（2011）

51. 张永昌、刘永、丁芳、斯晓明:扩散与竞争的稳定性研究
在线主题之间。Int. J. Mod. Phys. C 21(12), 1517–1529 (2010)
52. 赵建: 基于粒度计算的Web使用挖掘. 华南理工大学, 广州 (2010)

53.Zhao, K., Kumar, A.:谁写什么博客:了解博主的发布行为。全球资讯网。号码:10.1007/s11280-012-0167-3 54. 赵平,蔡庆生,王庆阳,Gen,HT:一种
基于复杂网络特征的中文文档关键词自动提取算法。模式识别与

人工智能20(6), 827–831 (2007)

55. Zhong, N., Bradshaw, JM, Liu, JM, Taylor, JG: 脑信息学。IEEE Intell. Syst. 26(5),
16–21（2011年）
56. 钟宁、马建华、黄瑞华、刘建明、姚洋、张永新、陈建华：智慧物联网（W2T)的挑战与展望研究。 J.超级计算机。 1-21
（2010）。 doi:10.1007/s11227-010-0518-8

57.周YD,孙QD,关XH,李W.,陶J.:流量内容词相关性的互联网热门话题提取。西安交通大学学报41(10), 1142–1145
(2007)

58. Zhu, MX, Cai, Z., Cai, QS: 使用小世界结构自动提取中文文档关键词。自然语言处理与知识工程学报,第 438-443
页（2003 年）

59. 朱涛:社交网络中节点角色与群体演化研究。北京邮电大学,北京（2011）