影响力扩散概率模型及其用于意见领袖发现研究

樊兴华[1)] 赵 静[1)] 方滨兴[2)] 李欲晓[3)]

[1)](重庆邮电大学中文信息处理研究所 计算智能重庆市重点实验室 重庆 400065)

[2)](北京邮电大学计算机学院 北京 100876)

[3)](北京邮电大学国际学院 北京 100876)

摘 要 作为意见领袖识别基础的影响力扩散模型 IDM 存在两个缺陷:(1)由回复链结构断层或者帖子内容间接传播引起的影响力传递中断;(2)由灌水所导致的虚假影响力传递.为解决上述问题,文中提出了一种新的影响力扩散概率模型 IDPM,进而建立了网络意见领袖筛选模型.该模型在相同兴趣空间上定义单个关键词语传播概率影响力,在帖子影响力定义中引入了有效关键词语概念,避免了上述缺陷;同时,在用户影响力计算时给每个帖子一个影响因子,用以整合其它有用信息,使模型具有开放性和包容性特点.在 2010 年 12 月到 2011 年 5 月网易社会新闻版块评论数据上的实验表明,文中方法是有效的,其平均精确率相对 IDM 模型提高了 59.8%.

关键词 意见领袖识别;IDM;影响力扩散概率模型

中图法分类号 TP391 **DOI 号** 10.3724/SP.J.1016.2013.00360

Influence Diffusion Probability Model and Utilizing It to Identify Network Opinion Leader

FAN Xing-Hua[1)] ZHAO Jing[1)] FANG Bin-Xing[2)] LI Yu-Xiao[3)]

[1)](*Chongqing Key Laboratory of Computational Intelligence, Chinese Information Processing Laboratory, Chongqing University of Posts and Telecommunications, Chongqing* 400065)

[2)](*School of Computer Science, Beijing University of Posts and Telecommunications, Beijing* 100876)

[3)](*International School, Beijing University of Posts and Telecommunications, Beijing* 100876)

Abstract There exist two defects in the influence diffusion model, which is a base for opinion leader identification. One is the influence diffusion break caused by the broken reply chain or indirect content diffusion, and the other is illusive influence diffusion caused by flooding posts. To solve the above problems, this paper presents a new Influence Diffusion Probability Model (IDPM), and then builts a network opinion leader identification model. In which, the diffusion probability influence of the single term is defined in the same interesting space, and the concept of valid term in post influence evaluation is introduced. An impact fact to be added to influence calculation, which integrates other useful information, and this leads the model is open and inclusive. The experiments in the data collected from the NetEase social news section from December 2010 to May 2011, show that the method proposed in this paper is valid, and the average precision is more 59.8% than that of IDM.

Keywords opinion leader identification; IDM; influence diffusion probability model

收稿日期:2011-12-05;最终修改稿收到日期:2012-05-29.本课题得到国家自然科学基金项目(60703010)和国家自然科学基金重大研究计划重点培育类项目"非常规突发事件中网络舆情作用机制与相关技术研究"(90924029)资助.樊兴华,男,1972 年生,博士,教授,主要研究领域为人工智能、自然语言处理、信息检索、网络内容安全、不确定性推理和复杂系统故障诊断. E-mail: fanxinghua1972@gmail.com. 赵 静,女,1986 年生,硕士研究生,主要研究方向为中文信息处理. 方滨兴,男,1960 年生,博士,教授,博士生导师,中国工程院院士,主要研究领域为信息安全等. 李欲晓,男,1968 年生,博士,教授,主要研究领域为信息传播、立法和公共安全.

1 引　言

随着互联网对人们生活的逐渐渗透，越来越多的人习惯于借助新兴媒体，如微博客、QQ、MSN、聊天室和论坛等平台交流心得体会，参与公众话题讨论. 在信息传播的过程中，意见领袖作为一种重要力量，在社会舆论的形成过程中发挥着不可忽视的作用，局部意见在意见领袖的引导下演化为舆论，影响力直接渗透到现实社会[1].

意见领袖(opinion leader)概念最早是由美国学者拉扎斯菲尔德等在《人民的选择》一书中提出，其认为观点和意见从媒体流向意见领袖，再从意见领袖流向人群中不太活跃的用户[2]. 本文的网络意见领袖是指那些通过在新兴媒体发表帖子(文本)、或者回复其他网络用户发表的帖子这种基于文本的交流方式，将自己的见解、观点传递给其他网络用户，引起他们内心的共鸣，进而影响、改变他们的观点、思想和决策的网络用户. 意见领袖发现任务就是从海量的具有回复关系的帖子中找出那些具有重要影响的网络用户(意见领袖). 本文的网络意见领袖和传统意见领袖在概念上有着本质的区别，他们的社会特征(例如学位、职位等信息)被隐藏，因此一些基于量表、问卷调查方式的传统意见领袖识别方法[3](例如自我报告法、网络分析方法等)不适用于本文的意见领袖发现.

随着新兴媒体的兴起，网络舆论影响力的不断加大，网络意见领袖研究也引起了国内外学者的积极关注，他们主要从 3 个方向进行研究：(1) 以帖子文本内容为侧重点，通过考虑帖子中词语在回复结构中的影响传播来识别意见领袖[4-6]. 该方向最具代表性的研究是日本学者松村直弘等提出的影响力扩散模型 IDM(Influence Diffusion Model)[4]，IDM 模型是该方向的基石，其它模型和方法都采用它来计算帖子的影响力. (2) 以帖子回复结构为侧重点，通过考虑根据帖子回复结构构建网络的网络特性来识别意见领袖[7-10]. (3) 以网络用户的统计信息(例如用户发表帖子总数，回复其他用户帖子总数，回复其他用户人数，回复该用户帖子的人数，回复该用户帖子的帖子总数等)为侧重点，利用统计信息建立模型来识别意见领袖[11-12]. 从网络意见领袖定义中可以看出，帖子内容是根本，它是引起其他网络用户内心共鸣的原因；帖子回复结构网络的网络特性和网络用户的统计信息是表象，它是由网络意见领袖和与之共鸣的网络用户构成整体所展示的外部现象. 因此，网络意见领袖发现应当以基于帖子内容的模型和方法为主，兼顾帖子结构网络的网络特性和网络用户统计信息. 反过来说，采用基于帖子回复结构网络特性和基于用户统计信息的模型和方法识别出的意见领袖精确性不高，其原因在于：包含意见领袖的帖子回复结构网络具有一定的网络特性，或者包含意见领袖的用户统计信息满足一定规律，它们是成立的；反之，它未必成立，这之间没有一一对应的关系. 例如文献[13]中采用帖子结构网络特性识别出的排名第一的意见领袖 Merits，尽管其发帖数目较大，是论坛的活跃分子，但从未得到过他人回复，没有意见领袖所拥有的影响力和号召力，不是真正的意见领袖. 可见，需要一种基于帖子内容的意见领袖发现模型，该模型应具有开放性和包容性，能够整合回复网络结构特性、用户统计信息，甚至将来发现的对意见领袖识别有用的其它信息的能力. 本文的一个目标就是建立这样一种模型和方法.

我们发现作为基于帖子内容意见领袖识别方向基石的 IDM 模型，存在影响力传递断层和由灌水导致的虚假影响力传播两个问题，致使以该模型为基础的意见领袖发现方法精度不高，阻碍了这一方向的发展和应用. 本文的另一目的就是研究一种更为合理的、可代替 IDM 模型的影响力传播模型，以此为基础建立意见领袖发现模型. 本文第 2 节分析 IDM 模型存在的缺陷；第 3 节建立意见领袖发现概率模型，它包括两部分：特征影响力传播概率模型和以此为基础的意见领袖筛选模型；第 4 节为实验和分析；最后总结全文内容.

2 影响力扩散模型(IDM)存在问题分析

2.1 IDM 模型的基本思想

IDM 模型通过挖掘蕴含在网络文本内容和回复结构中的规律来测量论坛参与者的活动，并假设论坛影响力最高的用户即为论坛意见领袖，它提出两个重要命题[4]：

(1) 在基于文本的计算机中介交流环境中，人们通过发帖、回帖表达观点，因此论坛回复链体现影响力的传递结构.

(2) 词语是组成帖子意义的基本单位，在基于文本的计算机中介交流环境中，论坛交流通过词语来表达和传播. 帖子影响力定义为帖子包含的词语集合在回复链传播的程度，采用回复关系的上下游

帖子的词语交集数与下游帖子词语数之比来进行计算.帖子回复链结构表示了个体之间的关系,一个个体的影响力就是他提交的所有帖子的影响力的总和.因此,通过帖子的影响力计算就可以找到最有影响力的个体,也就是意见领袖.

2.2 IDM 模型的缺陷

IDM 模型利用回复关系的上下游帖子的词语交集数与下游帖子词语数之比来度量帖子的影响力,在下述两种情况下与真实情况相悖.

(1)影响力传递断层现象,包括回复链结构中断和帖子内容传递中断两种情况.

回复链结构中断示例如图 1 所示,可简单地理解为用户 C 回复用户 B,但找不到用户 B 的根节点,即在一个话题树中出现悬浮的回复链.

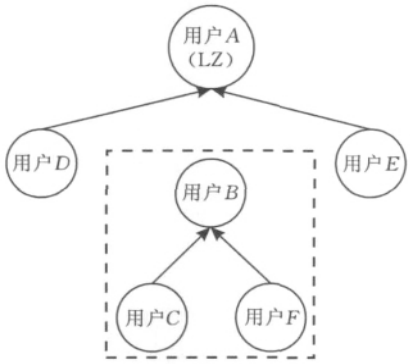

图 1 回复链结构中断示例

图 1 中用户 B、C、F 组成的回复链悬浮在以用户 A 为根节点的话题树中,根据 IDM 算法,用户 B 的影响力得分为 0,进而导致用户 B 对楼主 A 的影响力得分贡献为 0,减小了用户 A 成为意见领袖的可能.在我们实际搜集的语料中,以网易社会新闻版块的评论数据为例,共有 38 604 个用户参与话题讨论,其中有 22 058 个用户有跟帖,但找不到其根节点,即出现影响力传递断层现象.可见,这种回复链结构中断引起的影响力断层现象是很严重的.

帖子内容传递中断示例如图 2 所示.图中 C_3 帖对 C_1 帖进行回复,但没有传递 C_1 的关键词.C_5 帖又对 C_3 帖进行回复,其帖子内容不仅包含 C_1 帖的关键词 B,也包括 C_3 帖的关键词 F,即 C_1 帖通过 C_3 帖间接影响 C_5 帖,它们之间存在影响力传播.根据 IDM 模型算法,下游帖子 C_3 与上游帖子 C_1 的词语交集为空,则 C_1 帖对 C_3 帖的影响力为 0,而作为帖 C_3 下游帖子的 C_5 帖影响力也为 0.这是由于帖子 C_1 的关键词没有传到第 2 层的 C_3 帖,导致 C_1 帖的影响力传不到 C_5 帖,即下游帖子的影响力传播出现了断层.这种情况在实际的论坛、微博客等平台中是比较常见的.

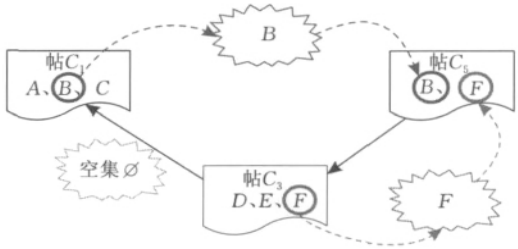

图 2 帖子内容传递中断示例

(2)虚假影响力传递现象.在实际的微博客、论坛等平台中,灌水现象是很常见的,这必然导致虚假影响力传递.例如,用户 A 在短时间内对另一个用户 B 进行重复回复,根据 IDM 模型算法,用户 B 的影响力得分就会很高,而据此选出的意见领袖显然不是真正的意见领袖,这与实际情况不符.

2011-04-24 08:23:02 网易湖南省手机网友 211.138.*.* ＊郑凯_杨苏剑 警察是有太才了,你家人自杀还脱衣服
2011-04-24 08:23:09 网易湖南省手机网友 211.138.*.* ＊郑凯_杨苏剑 警察是有太才了,你家人自杀还脱衣服
2011-04-24 08:23:13 网易湖南省手机网友 211.138.*.* ＊郑凯_杨苏剑 警察是有太才了,你家人自杀还脱衣服
2011-04-24 08:23:15 网易湖南省手机网友 211.138.*.* ＊郑凯_杨苏剑 警察是有太才了,你家人自杀还脱衣服

上面所列为我们实际搜集的网易社会新闻版块评论数据中的 4 条帖子,可以看到,用户“网易湖南省手机网友 211.138.*.*”在短时间内对楼主“郑凯_杨苏剑”进行重复回复,用户“网易湖南省手机网友 211.138.*.*”发的帖子中“警察、才、自杀、脱、衣服”均通过楼主“郑凯_杨苏剑”传递下来,经过影响力的累加,用户“郑凯_杨苏剑”的影响力得分就会很高.经统计,在该话题树中,这种灌水帖占整个帖子总数的 9.5%.

3 网络意见领袖发现概率模型

模型思路:认为在相同兴趣空间中,通过发帖、回帖这种基于文本交互的观点传递方式,影响力大的网络用户则为意见领袖.这样可将意见领袖的发现问题转化为求在相同兴趣空间中每个网络用户的影响力大小并排序.为避免出现类似 IDM 模型的缺陷,并使模型具有包容性和开放性,将模型分为建立

在帖子影响力估计上的影响力概率扩散模型，和在用户影响力估计上的意见领筛选模型. 前者通过在整个兴趣空间上定义单个关键词语传播概率影响力来解决 IDM 模型中的影响力传递结构断层问题和灌水导致的虚假影响力传播问题，通过考虑句子中的有效关键词语来解决 IDM 模型中的影响力传递内容断层问题；后者在计算用户影响力时，通过给每个帖子一个相应的影响因子，来整合模型的回复结构网络特性、用户统计信息、词语倾向性等可用信息，从而使模型具有开放性和包容性.

3.1　影响力扩散概率模型(Influence Diffusion Probability Model，IDPM)

用户帖子回复结构(简称 URS)定义为：由用户发帖导出的、通过回复关系链接在一起的帖子及其回复关系的集合. 在没有第 2 节所述的回复链结构中断的情况下，UerserRS 可用一个回复结构树(RST)来表示，树中的每个节点表示某用户的发帖(包含用户 ID、帖子文本、发帖时间等属性)，节点间的连接边表示回复关系. 在有回复结构中断情况下，UerserRS 可表示为多个回复结构树所构成的森林. 用户帖子回复结构分为主帖回复结构和跟帖回复结构两种，前者是由用户所发主帖导出的，它表示一个话题；后者是由用户所发跟帖导出的，它是从主帖回复结构中截取的一部分.

图 3 中，用户 ID_0 发表了主帖 A、B、C，用户 ID_1 和 ID_2 分别回复了帖子 A、C、D、E 和帖子 B、C、E、F，用户 ID_3 回复 ID_1 帖子 A、E、F，它们形成了一棵回复结构树.

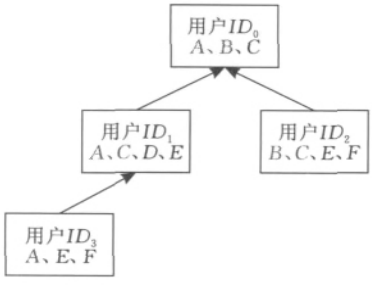

图 3　用户帖子回复结构示例

相同兴趣空间：由相关的多个主帖回复结构构成的集合. 这里的相关是指多个话题在内容上相近、类似或者关联度比较高. 相同兴趣空间可用主帖回复结构森林 G 来表示，它是该空间内所有主帖回复结构的集合.

相同兴趣空间内的关键词语：将相同兴趣空间中的所有文本内容看作一个整体，其中每个帖子看做一个文本，以词语作为特征，采用诸如文本分类[13]处理中使用的特征选择技术进行处理后得到特征，也就是关键词语. 每个帖子可用关键词语向量表示，帖子的内容按照回复结构进行传递. 引入关键词语可以有效地去除意见领袖发现中的噪声.

关键词语 $term$ 在主帖回复结构中的传播频率 $f_{\mathrm{URS}}(term)$ 定义：$term$ 在主帖回复结构中沿着回复结构传播的次数，$term$ 在上游帖子和回复它的下游帖子中同时出现，则表示 $term$ 被传播 1 次.

例如图 3 中，假设 A、B、C、D、E 和 F 均为关键词语，则它们的传播频率为

$f(A)=2$，A 从 ID_0 到 ID_1 传播了 1 次，从 ID_1 到 ID_3 传播了 1 次；

$f(B)=1$，B 从 ID_0 到 ID_2 传播了 1 次；

$f(C)=2$，C 从 ID_0 到 ID_1 传播了 1 次，从 ID_0 到 ID_2 传播了 1 次；

$f(D)=0$，D 没有被传播；

$f(E)=1$，E 从 ID_1 到 ID_3 传播了 1 次；

$f(F)=0$，F 没有被传播.

关键词语 $term$ 在相同兴趣空间 G 内传播频率 $f_G(term)$ 定义：$term$ 在相同兴趣空间内所有主帖回复结构中传播频率之和：

$$f_G(term)=\sum_{i=1}^{|G|} f_{URS_i}(term) \tag{1}$$

其中 $|G|$ 表示相同兴趣空间 G 内所包含的主帖回复结构数目.

关键词语 $term$ 在相同兴趣空间 G 内的影响传播概率 $P_{\mathrm{idp}}(term)$ 定义为

$$P_{\mathrm{idp}}(term)=\frac{f_G(term)}{\sum_{j=1}^{m} f_G(term_j)} \tag{2}$$

其中 m 为相同兴趣空间 G 内所包含关键词语数目.

用户所发帖子 X 的影响力 INF_X 定义：帖子中被有效传递关键词语的影响传播之和. 假设帖子的关键词语向量为 $\langle term_1, term_2, \cdots, term_n\rangle$，$INF$ 采用如下公式计算：

$$INF_X=\sum_{i=1}^{n} w_i \lg P_{\mathrm{idp}}(term_i) \tag{3}$$

其中 w_i 为用户所发帖子 X 的有效传递因子. 帖子关键词语 $term$ 的所谓有效传播是指它在以该帖子 X 为根的帖子回复结构中至少再出现一次. w_i 定义如下：

$$w_i=\begin{cases}1, & term_i \text{为有效关键词语}\\ 0, & \text{其它}\end{cases}.$$

3.2 意见领袖筛选模型

一个网络用户可能发表多个帖子，这些帖子中既有主帖，也有对其他用户的回复帖，甚至还有一些灌水帖. 有些帖子可能受到大家的热捧，有些帖子观点也会受到其他用户的批判. 除去基于传播内容的影响之外，帖子的其它因素也可能会对意见领袖识别有贡献，意见领袖筛选模型应当具有开放性、包容性和可扩展性，能够考虑这些因素.

用户的影响力 INF_{User} 定义为：用户所发全部有效帖子的影响力之和. 假设用户发表了 m 个帖子，INF_{User} 计算公式如下.

$$INF_{\text{User}}=\sum_{i=1}^{m}u_i INF(X_i) \tag{4}$$

其中，u_i 为帖子有效系数，它用于整合诸如帖子回复结构网络特性、用户统计信息、帖子倾向性等因素对用户最终影响力计算的影响. 由于本文重点在于如何解决 IDM 模型存在的问题，进而提出替代模型. 因此，本文实验中 $u_i\equiv 1$，如何整合其它信息的问题将另文专门讨论.

3.3 本文模型特点分析

和 IDM 模型一样，本文提出的 IDPM 模型也是以在回复结构中传递的帖子词语为基础来度量帖子的影响力. IDM 模型直接利用上下游两个帖子中共同出现词语数与下游帖子的词语数比值来度量影响力，这导致了第 2 节分析出的几个缺陷. IDPM 模型将影响力的度量分成了两部分，一是式(2)中蕴含的关键词语 $term$ 在相同兴趣空间 G 内的影响传播概率，它直接定义在整个兴趣空间的回复结构之上，而不是像 IDM 那样定义在上下游两个帖子的回复结构上，这自然就能够避免 IDM 模型中出现的结构中断问题；二是式(3)中蕴含的用户所发帖子 X 影响力，在它的定义中引入了有效传递关键词语的概念，这里的有效传递关键词语不仅包含上下游两个帖子中直接传播的词语，也包含了通过下游帖子间接传递的词语，这样自然就避免了 IDM 模型中出现的内容中断问题. IDPM 模型将词语 $term$ 在整个兴趣空间的出现作为一个整体看待，这通过式(1)中蕴含的关键词语在相同兴趣空间 G 内传播频率来实现，这种处理稀释了灌水帖子的影响，抑制了灌水帖导致的虚假影响力传播. 建立在 IDMP 模型上的由式(4)描述的意见领袖筛选模型，由于引入了帖子有效系数，用于整合其它各种可用信息，具有开放性、包容性特点. 可见本文提出模型在理论上就能克服 IDM 模型存在的缺陷，具有优越性.

4 实 验

4.1 实验数据

将门户网站网易的社会新闻版块看作相同兴趣空间，利用网易内嵌的搜索引擎有道，在高级搜索中限制时间、新闻类别、排序方式以及新闻源，借助 MetaStudio 和 DataScraper 抓取工具，收集了该板块从 2010 年 12 月到 2011 年 5 月的评论数据作为实验数据集.

该数据集包含 1170 个话题(去除了参与人数过少的话题)，共有 116213 个帖子，38604 个参与讨论的用户. 其中，十万级参与人的话题 33 个，上万级参与人的话题 126 个，上千级参与人的话题 247 个，上百级参与人的话题 525 个. 单个话题树的最大深度为 100.

4.2 评估指标

迄今为止，还没有一套大家公认的、统一的意见领袖识别处理评估方法. 本文中采用类似信息检索的评估方法[14]来进行评价. 因为意见领袖识别可看成一种类似的信息检索处理，即它从相同兴趣空间内、以回复结构链接起来的帖子数据中，以用户的影响力作为查询进行的一种检索处理. 采用了如下指标：

前 N 个结果中的正确率 $P@N$：

$$P@N=\frac{\text{前 }N\text{ 个结果中人工判定为真正意见领袖的个数}}{N}.$$

前 N 个结果的平均正确率 $AvgP@N$：

$$AvgP@N=\sum_{i=1}^{N}P@i/N.$$

实验中，两名工作人员根据用户所发全部帖子内容以及用户的活跃度、受关注度和影响力覆盖度这 3 个统计指标来判定某一用户是否为一名真正的意见领袖.

4.3 实验方法

为了验证、评估本文方法，我们对比了如下 3 种方法：

(1) OL_IDPM，本文提出的意见领袖发现概率模型方法.

(2) OL_IDM，基于影响力传播模型 IDM 的意见领袖发现方法.

(3) OL_IDM_P，基于改进影响力传播模型 IDM_P 的意见领袖发现方法.

通过随 N 变化的 $P@N$ 曲线图来定性地评估不同方法的效果，根据 N 取不同值时的 $AvgP@N$ 来定量地评估不同方法的效果.

改进的影响力传播模型 IDM_P 认为[6]，下游帖子中不被传播的关键词的多少不应该影响上游帖子对其的影响，IDM 模型公式中的分母项对影响力的计算没有实际意义，考虑通过省去公式中的分母项来改进 IDM 模型. 该模型也是从上下帖子中共同出现的词语来度量影响力的传递，与 IDM 模型本质上是一样的，存在第 2 节描述的缺陷. 作为一种 IDM 模型的改进版本，本文将它作为一种参照，用于评估本文方法.

4.4　实验结果及分析

表 1 给出了上述 3 种方法的对比实验结果，为了节约篇幅，在下面的实验中只给出排名前 5 的用户.

表 1　3 种模型对比实验结果(前 5 名用户)

排名	OL_IDM		OL_IDM_P		OL_IDPM	
	UserID	Score	UserID	Score	UserID	Score
1	郑凯_杨苏剑	1063	郑凯_杨苏剑	8431	刘薇	−153827
2	魏绘轩_王振东	306	褚朝新_郭少峰_邢世伟	5222	马喜生	−134807
3	曹晶晶_余亚莲	301	曹晶晶_余亚莲	4829	曹晶晶_余亚莲	−130780
4	洪雪	299	马喜生	4561	褚朝新_郭少峰_邢世伟	−119791
5	张鹏翔_李铁锤	295	张鹏翔_李铁锤	4094	郑凯_杨苏剑	−90215

从表 1 来看，OL_IDPM 与 OL_IDM 和 OL_IDM_P 的排名差异主要在用户“郑凯_杨苏剑”和“刘薇”上. 观察实际语料，可知以用户“郑凯_杨苏剑”为根节点的话题树中，有 9.5%的帖子属于重复帖，即对用户“郑凯_杨苏剑”进行灌水式的回复，由此得出的意见领袖显然不属于真正的意见领袖. 而 OL_IDPM 以关键词语在整个内容空间的出现来实现，在一定程度上稀释了灌水帖的影响，相应的“郑凯_杨苏剑”的排名较靠后. 以用户“刘薇”为根节点的话题树中，共有 544 个节点，其中楼主为第 1 层节点，第 2 层节点有 365 个，其它的 178 个节点为 3 层及以下的. 经过统计分析，3 层及以下的 178 个节点中，有近 3/4 的用户无法通过中间节点挂到楼主“刘薇”节点上，即帖子结构出现断层，导致在 OL_IDM 和 OL_IDM_P 中，均没能排到前 5. OL_IDPM 考虑整个兴趣空间中的回复结构，有效地解决了 OL_IDM 和 OL_IDM_P 的结构断层问题.

表 2 是前 50 名用户中 3 种方法筛选出的真正意见领袖所占的比例. 表 2 中，OL_IDPM 在解决影响力断层和灌水帖问题的基础上，提高了 OL_IDM 和 OL_IDM_P 发现真正意见领袖的比例，较之最初的 OL_IDM 方法，其前 50 名的正确率提高了 71.4%.

表 2　前 50 名用户对比结果

方法	判断正确/个	相比 OL_IDM 改变率/%	相比 OL_IDM_P 改变率/%
OL_IDM	21	—	—
OL_IDM_P	31	+47.6	—
OL_IDPM	36	+71.4	+16.1

综合表 1 和表 2，可见本文所提模型能够有效地解决 IDM 模型存在的问题.

为了定性评估本文 OL_IDPM 的有效性，下面给出前 100 名用户的正确率变化曲线图，如图 4 所示.

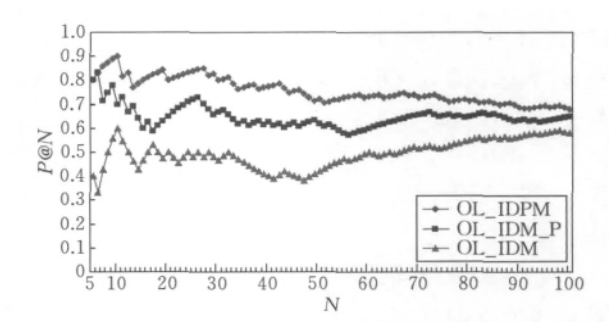

图 4　前 100 个结果的正确率变化曲线

图 4 中，比较 OL_IDPM、OL_IDM 和 OL_IDM_P 可以看出，OL_IDPM 明显优于前两个模型，尤其是 $P@5$ 的值，即概率模型在解决了 OL_IDM、OL_IDM_P 模型所存在问题的同时，提高了发现真正意见领袖的准确率.

表 3～表 7 分别是 N 取 5、10、20、50、100 时的平均正确率 $AvgP@N$，以便我们定量地评估 OL_IDPM、OL_IDM 和 OL_IDM_P 这 3 种方法.

表 3　*AvgP@5* 的对比结果

方法	平均正确率/%	相对 OL_IDM 的改变率/%	相对 OL_IDM_P 的改变率/%
OL_IDM	19.6	—	—
OL_IDM_P	54.4	+177.6	—
OL_IDPM	96.0	+389.8	+76.5

表 4 *AvgP*@10 的对比结果

方法	平均正确率/%	相对 OL_IDM 的改变率/%	相对 OL_IDM_P 的改变率/%
OL_IDM	34.1	—	—
OL_IDM_P	65.0	+90.6	—
OL_IDPM	91.5	+168.3	+40.8

表 5 *AvgP*@20 的对比结果

方法	平均正确率/%	相对 OL_IDM 的改变率/%	相对 OL_IDM_P 的改变率/%
OL_IDM	41.5	—	—
OL_IDM_P	64.7	+55.9	—
OL_IDPM	86.4	+108.2	+33.5

表 6 *AvgP*@50 的对比结果

方法	平均正确率/%	相对 OL_IDM 的改变率/%	相对 OL_IDM_P 的改变率%
OL_IDM	43.3	—	—
OL_IDM_P	64.7	+49.4	—
OL_IDPM	81.7	+88.7	+26.3

表 7 *AvgP*@100 的对比结果

方法	平均正确率/%	相对 OL_IDM 的改变率/%	相对 OL_IDM_P 的改变率/%
OL_IDM	48.0	—	—
OL_IDM_P	64.0	+33.3	—
OL_IDPM	76.7	+59.8	+19.8

从表 3～表 7 可以看出，方法 OL_IDM_P 优于方法 OL_IDM，方法 OL_IDPM 优于方法 OL_IDM_P，且显著优于方法 OL_IDM，其 *AvgP*@100 分别提高了 19.8%和 59.8%.

5 结 论

如何有效地筛选出真正的网络意见领袖是一项具有挑战性的现实任务. 作为意见领袖识别基础的影响力扩散模型 IDM 存在两个缺陷：(1) 由回复链结构断层或者帖子内容间接传播引起的影响力传递中断；(2) 由灌水所导致的虚假影响力传递. 本文提出了一种新的影响力扩散概率模型 IDPM，在此基础上建立了网络意见领袖筛选模型. 该模型从理论上避免 IDM 模型的缺陷，具有开放性和包容性特点. 实验验证了本文方法的有效性，明显优于 IDM 模型及其改进模型.

在下一步的工作中，我们将考虑如何利用帖子有效系数，整合帖子回复结构网络特性、用户统计信息、帖子倾向性等因素，以满足筛选出不同网络意见领袖定义的需求.

参 考 文 献

[1] Liu Jian-Ming. Public Opinion Propagation. Beijing: Tsinghua University Press, 2001(in Chinese)
(刘建明. 舆论传播. 北京：清华大学出版社，2001)

[2] Lazarsfield P et al. The People's Choice. New York: Columbia University Press, 1948

[3] Weimann Gabriel, Tustin Deon Harold, van Vuuren Daan, Joubert J P R. Looking for opinion leaders: Traditional vs. modern measures in traditional societies. International Journal of Public Opinion Research, 2007, 19(2): 173-190

[4] Matsumura Naohiro, Ohsawa Yukio, Ishizuka Mitsuru. Influence diffusion model in text-based communication. Transactions of the Japanese Society for Artificial Intelligence, 2002, 17(3): 259-267

[5] Yu Hong. Research on the opinion leaders of political BBS: An case study on Sino-Japan BBS of strong nation forum [Ph. D. dissertation]. Huazhong University of Science and Technology, Wuhan, 2007: 32-36(in Chinese)
(余红. 网络时政论坛舆论领袖模型初探[博士学位论文]. 华中科技大学，武汉，2007：32-36)

[6] Shi Mao, Fang Yong, Zeng Xiang-Ping, Wang Chang-Hui. Analysis of IDM model and its reformative arithmetic. Journal of Chengdu University of Information Technology, 2008, 23(1): 69-72(in Chinese)
(石矛，方勇，曾祥平，王长辉. IDM 模型分析及其影响力改进算法. 成都信息工程学院学报，2008，23(1)：69-72)

[7] Zhou Hengmin, Zeng Daniel, Zhang Changli. Finding leaders from opinion networks//Proceedings of the 2009 IEEE International Conference on Intelligence and Security Informatics (ISI'09). Dallas TX, USA, 2009: 266-268

[8] Gao Jun-Bo, Yang Jing. Analysis of opinion leader in on-line communities. Journal of University of Electronic Science and Technology of China, 2007, 36(6): 1249-1252(in Chinese)
(高俊波，杨静. 在线论坛的意见领袖分析. 电子科技大学学报，2007，36(6)：1249-1252)

[9] Zhang Jun, Ackerman Mark S, Adamic Lada. Expertise networks in online communities: Structure and algorithms//Proceedings of the 16th International World Wide Web Conference Committee (IW3C2). Banff, Canada, 2007: 221-230

[10] Song Xiaodan, Chi Yun, Hino Koji, Belle Tseng. Identifying opinion leaders in the blogosphere//Proceedings of the 16th ACM Conference on Information and Knowledge Management (CIKM'07). New York, USA, 2007: 971-974

[11] Zhai Zhongwu, Xu Hua, Jia Peifa. Identifying opinion leaders in BBS//Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT'08). Sydney, NSW, Australia, 2008: 398-401

[12] Wang Jue, Zeng Jian-Ping, Zhou Bao-Hua, Wu Cheng-Rong. Online forum opinion leaders discovering method based on clustering analysis. Computer Engineering, 2011, 37(5): 44-46(in Chinese)
(王珏，曾剑平，周葆华，吴承荣. 基于聚类分析的网络论坛意见领袖发现方法. 计算机工程，2011，37(5)：44-46)

[13] Fan Xing-Hua, Sun Mao-Song. A high performance two-class Chinese text categorization method. Chinese Journal of Computers, 2006, 29(1): 124-131(in Chinese)
(樊兴华，孙茂松. 一种高性能的两类中文文本分类方法. 计算机学报, 2006, 29(1): 124-131)
[14] Fan Xing-Hua, Nie Jian-Yun. Link distribution dependency model for document retrieval. Journal of Information & Computational Science, 2009, 6(3): 1553-1564

FAN Xing-Hua, born in 1972, Ph.D., professor. His research interests include artificial intelligence, natural language processing, information retrieval, network content security, and uncertain reasoning and fault diagnosis to complex system.

ZHAO Jing, born in 1986, M.S. candidate. Her research interests focus on natural language processing.

FANG Bin-Xing, born in 1960, Ph.D., professor, Ph.D. supervisor, member of Chinese Academy of Engineering. His current research interests include information security etc.

LI Yu-Xiao, born in 1968, Ph. D., professor. His research interests include dissemination of information, Legislation and public policy.

Background

This paper is related to the mechanism of public opinion on internet for abnormal emergency. Because abnormal emergency owns the characteristics of explosive, uncertain evolution and group diffusion, more and more researchers concentrate on and try to solve it. Especially, as the carrier and platform of broadcasting of abnormal emergency, internet along with its popularity makes public opinion on internet absorb more attention. Some scholars have already done some research, but most of them focus on the qualitative research of the evolution of public opinion. It is a new viewpoint to research on the development of abnormal emergency from the qualitative and quantitative analysis.

In this paper, we propose an Influence Diffusion Probability Model and then build a network opinion leader identification model. In which, we define the diffusion probability influence of the single term in the same interesting space, and the concept of valid term in post influence evaluation is introduced. An impact fact to be added to influence calculation, which integrates other useful information, and this leads the model is open and inclusive. This paper presents several experiments to support our model, in which a large number of real forum data collected from the NetEase social news section are used. The experiments show that method proposed in our paper is valid.

This work is a part of the "Influence Diffusion Probability Model and Utilizing It to Identify Network Opinion Leader", which is mainly supported by the National Natural Science Foundation of China (Grant No. 60703010) and the Major Research Plan of the National Natural Science Foundation of China (Grant No. 90924029).