有这个性质,则称 X_j 为 o^* 的一个最小改进变量.最小改进变量是指在 N 中可通过翻转得到优于配置 o^* 的最小变量.当且仅当 X_j 不属于与 o 后缀匹配变量中的最小改进变量时,我们称 X_j 为目标配置 o 的最小改进变量.换言之,如果 o 和 o^* 的某些后缀匹配,则根据最小改进变量的定义,需将注意那些不属于后缀匹配的变量.

最小变量翻转规则不区分不同最小改进变量的翻转,只是简单地限制翻转.一般来说,并非所有的最小翻转变量都适合,有些可能导致死点,从而需要进行回溯.当我们回溯时,我们只需考虑其他最小变量翻转,而不是所有的翻转,从而显著减少搜索树的大小以及预期的回溯量.我们注意到,二值树结构 CP-net 的 TreeDT 算法本质上实现了最小变量翻转规则,因此,这是一个对二值树结构网络的完整并且无回溯的搜索过程的算法.

例 3.1 给定包含三个变量 A 、B 、C 的 CP-net,其中 $\mathrm{pa}(B)=A$, $\mathrm{pa}(C)=B$, $\mathrm{dom}(A)=\{a,\bar{a}\}$, $\mathrm{dom}(B)=\{b_1,b_2,b_3\}$, $\mathrm{dom}(C)=\{c,\bar{c}\}$. 给定以下条件偏好:

$a\succ\bar{a}$;

$a:b_3\succ b_2\succ b_1$;

$\bar{a}:b_3\succ b_1\succ b_2$;

$b_1:c\succ\bar{c}$ $b_2:\bar{c}\succ c$ $b_3:c\succ\bar{c}$

给定查询 $N\models a\,b_3\,\bar{c}\succ\bar{a}\,b_1c$. 在使用 DFS 算法的情况下,需先得出该 CP-nets 的导出图,从而需要对 12 个配置进行 66 次比较.根据最小翻转变量规则可知,c 在 b_1 的情况下不能改进,但 b_1 在 $\bar{a}$ 的情况下可以改进到 b_3. 事实上,这是配置 $\bar{a}\,b_1c$ 唯一的一个最小变量翻转.这个翻转导向配置 $\bar{a}\,b_3c$,但该配置与目标配置 $a\,b_3\,\bar{c}$ 之间不存在路径.相反,先把非最小改进变量 A 翻转为 a 可得到一个成功的改进路径:先把配置 $\bar{a}\,b_1c$ 从 $\bar{a}$ 翻转到 a 后得到 $a\,b_1c$,有 $a\,b_1c\succ\bar{a}\,b_1c$;然后把 b_1 翻转成 b_2 得到 $a\,b_2c$,有 $a\,b_2c\succ a\,b_1c$;再将 c 翻转成 $\bar{c}$ 得到 $a\,b_2\,\bar{c}$,有 $a\,b_2\,\bar{c}\succ a\,b_2c$;最终把 b_2 翻转为 b_3 得到目标配置 $a\,b_3\,\bar{c}$,有 $a\,b_3\,\bar{c}\succ a\,b_2\,\bar{c}$. 根据偏好的传递性有 $a\,b_3\,\bar{c}\succ a\,b_2\,\bar{c}\succ a\,b_2c\succ a\,b_1c\succ\bar{a}\,b_1c$,从而得出满足用户偏好的查询 $N\models a\,b_3\,\bar{c}\succ\bar{a}\,b_1c$.通过使用最小翻转变量规则,我们发现,为得出满足用户偏好的查询 $N\models a\,b_3\,\bar{c}\succ\bar{a}\,b_1c$,需进行 4 次比较.与使用 DFS 算法 66 次比较相比,最小翻转变量可有效减少参与比较的配置数量,进而减少配置间的比较次数,提高了查询效率.

3.3 向前修剪技术

在改进翻转序列的搜索过程中,都可使用通用的向前修剪技术.这种技术具有如下性质:

(Ⅰ)它通常可以很快表明没有翻转序列是可能的;

(Ⅱ)它通过修剪变量的域以减少翻转搜索空间;

(Ⅲ)它不影响健全性或完整性;

(Ⅳ)成本相对较低:其时间复杂度为 $O(nrd^2)$,其中 n 是变量的数量,r 是每个变量的条件偏好规则的最大数量,d 是最大可变域的大小.

一般是通过向前扫描,修剪任何不能出现在给定查询的改进翻转序列中的变量的值.我们考虑一组翻转,可能会忽略其父亲的相互依赖关系以及父亲可以改变他们的值的次数.

本文以与网络拓扑一致的顺序考虑变量(使节点的父节点先于节点之前考虑).对于每个变量 X_j 与父亲 U,我们构建一个域转换图,其中节点对应可能的值 $x_i\in\mathrm{dom}(X_j)$. 对于每个条件偏好关系 $\succ_{j_u}$ 在 $\mathrm{dom}(X_j)$ 上的形式:

$$xu_1\succ xu_2\succ\cdots\succ xu_d,$$

使得 u 仅包含 X_j 的父节点 U 的未经修改的值,我们在排序 $\succ_{j_u}$ 中的连续值之间包括有向弧(即对每个 $1<i\leqslant d$,从 x_i 到 x_{i-1} 的弧).

当回答查询 $N\models o\succ o'$ 时,我们可以在 X_j 的域转换图中修剪不在有向路径上的任何 X_j 的值,从 $o'[X_j]$ 到 $o[X_j]$. 这可以通过运行著名的 Dijkstra 算法[14]两次来实现:一次找到从 $o'[X_j]$ 到达的节点,再次找到可以到达 $o[X_j]$ 的节点.将这些节点集合相交,以便在 $T(o')$ 中从 $o'[X_j]$ 到 $o[X_j]$ 的任何路径找到 X_j 的可能值(即沿着从 $o'[X_j]$ 到 $o[X_j]$ 的任何改进序列).如果交叉点为空,则占优查询失败:从 $o'[X_j]$ 到 $o[X_j]$ 没有翻转序列.

例 3.2 给定 CP-net N 以及变量 $A,B,\cdots$,对于变量 A,有 $a\succ\bar{a}$,对于变量 B,有 $b\succ\bar{b}$. 对于查询 $N\models a\,\bar{b}\cdots\succ\bar{a}b\cdots$,我们首先考虑变量 A. A 的域转换图中包含 $\bar{a}\to a$,因此不需修剪 A 的值.如果 A 有第三个变量 $\bar{\bar{a}}$,并且有 $a\succ\bar{a}\succ\bar{\bar{a}}$,那么可以修剪 A 的第三个值,从而简化 A 的所有子代的条件偏好表.

对于变量 B，其域转换图中包含 $\bar{b} \rightarrow b$. 由于在 B 的域转换图中，查询 $\bar{b}$ 的更优选值中 B 的值不能从查询 b 的较不优选值中 B 的值到达，因此导致查询失败而不需查询其他变量.

如果使用 DFS 算法进行处理，则需根据偏好遍历该 CP-nets 导出图中的路径上的配置，直到所有配置比较结束才知是否存在满足用户偏好的查询需求；而使用向前修剪技术，则可根据查询需求快速发现变量的值是否存在可翻转序列，若不存在，则查询失败而不需继续查询，从而提高查询效率。若存在，则可对变量的域转换图进行修剪，减少不必要的变量值，从而减少搜索空间，提高查询效率.综上，通过向前修剪技术的使用，可有效减少搜索空间，进而提高查询效率.

通过例 3.2 发现，在不影响搜索过程正确性或完整性的前提下，通过使用剪枝技术对搜索树进行修剪均得出满足用户查询需求的一条路径，从而其时间复杂度为线性时间，即 $O(1)$. 由此可知，与 DFS 算法的多条路径相比，剪枝技术可有效减少搜索空间，从而提高查询效率.

4　结论

本文主要利用剪枝技术对占优查询进行简化，减少搜索空间，进而提高查询的效率.同时，本文也较为全面地论述了 CP-nets 的基本概念，为 CP-nets 的基础理论研究和应用奠定了基础.进一步研究方向为利用帕累托复合技术及其规则对偏好关系不确定的配置进行调整，从而保证可以有效地进行占优查询.

参考文献(References)

[1] LIU J, LIAO S. ExpressiveEfficiency of Two Kinds of Specific CP-Nets[M]. Elsevier Science Inc., 2015.

[2] LIU J L. Research on CP-nets and its expressive power [J]. Acta Automatica Sinica, 2011, 37(3): 290-302.

[3] BRAFMAN R I, DOMSHLAK C. Preference handling-an introductory tutorial [J]. AI Magazine, 2009, 30(1): 58-86.

[4] BOUTILIER C, BRAFMAN R I, DOMSHLAK C, et al. CP-nets: A tool for representing and reasoning with conditional ceteris paribus preference statements [J]. Journal of Artificial Intelligence Research, 2011, 21(1): 135-191.

[5] ZANUTTINI B. Learning conditional preference networks with queries [C]// International Joint Conference on Artificial Intelligence. Morgan Kaufmann Publishers Inc. 2009: 1930-1935.

[6] ZANUTTINI B. Learning conditional preference networks[J]. Artificial Intelligence, 2010, 174(11): 685-703.

[7] CONITZER V. Making decisions based on the preferences of multiple agents[J]. Communications of the ACM, 2010, 53(3): 84-94.

[8] TANG P Z, LIN F Z. Computer-aided proofs of Arrow's and other impossibility theorems [J]. Artificial Intelligence, 2009, 173(11): 1041-1053.

[9] LANG J. Logical preference representation and combinatorial vote [J]. Annals of Mathematics & Artificial Intelligence, 2004, 42(1-3): 37-71.

[10] BRAFMAN R I, DOMSHLAK C, SHIMONY S E. On graphical modeling of preference and importance [J]. Journal of Artificial Intelligence Research, 2006, 25(1): 389-424.

[11] BOUVERET S, ENDRISS U, LANG J. Conditional importance networks: a graphical language for representing ordinal, monotonic preferences over sets of goods[C]// Proceedings of the 21st International Joint Conference on Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers Inc. 2009: 67-72.

[12] 孙雪姣，刘惊雷. CP-nets 的可满足性及一致性研究[J]. 计算机研究与发展，2012，49(4)：754-762.
SUN X, LIU J L. On the satisfiability and consistency for CP-nets [J]. Journal of Computer Research & Development, 2012, 49(4):754-762.

[13] 刘惊雷，廖士中，张伟. CP-nets 的完备性及一致性研究[J]. 软件学报，2012，23(6):1531-1541.
LIU J L, LIAO S Z, ZHANG W. On the completeness and consistency for CP-nets[J]. Journal of Software, 2012, 23(6):1531-1541.

[14] CORMEN T T, LEISERSON C E, RIVEST R L. Introduction to algorithms [J]. Resonance, 1996, 1(9): 14-24.

第48卷第9期 中国科学技术大学学报 Vol.48,No.9
2018年9月 JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA Sep. 2018

文章编号:0253-2778(2018)09-0730-09

多关系社交网络中基于兴趣匹配的网络舆情传播模型

孙更新,宾 晟

(青岛大学数据科学与软件工程学院,青岛 266071)

摘要:以网络爬虫方式获取新浪微博用户属性信息及微博内容数据,利用数据挖掘技术从中发现微博用户间的多种显式和隐式关系.在此基础上,提出一种基于半监督学习的用户兴趣匹配预测算法,参照仓室模型的传播个体状态划分方法,基于传播个体间的兴趣匹配度界定各状态之间的转移过程和转移概率,进而构建基于用户兴趣匹配的网络舆情传播模型.研究结果表明,该模型能够较好地描述社交网络中的舆情传播规律,重现网络舆情在社交网络中的真实传播过程链.

关键词:舆情传播;传播模型;复杂网络;数据挖掘;社交网络

中图分类号:TP391 **文献标识码**:A doi:10.3969/j.issn.0253-2778.2018.09.007

引用格式:孙更新,宾晟. 多关系社交网络中基于兴趣匹配的网络舆情传播模型[J]. 中国科学技术大学学报,2018,48(9):730-738.

SUN Gengxin, BIN Sheng. Network public opinion propagation model based on interest matching in multiple relationship social network[J]. Journal of University of Science and Technology of China, 2018,48(9):730-738.

Network public opinion propagation model based on interest matching in multiple relationship social network

SUN Gengxin, BIN Sheng

(*College of Data Science and Software Engineering, Qingdao University, Qingdao* 266071)

Abstract: The relationship between user profiles and the data of microblog content in Sina microblog was obtained by programming and web crawler, and a variety of explicit or implicit relationships between microblog users were discovered by using data mining. On the basis of this, a semi-supervised user interest matching prediction algorithm was proposed. According to the individual state division method of compartment model, a network public opinion propagation model is constructed based on user interest matching through state transition analysis and inference of state transition probability. The results show that the model can well describe the laws of public opinion propagation in social networks, and reproduce the real propagation process of network public opinion in the social network from the perspective of complex networks.

Key words: public opinion propagation; propagation model; complex network; data mining; social networks

收稿日期:2018-05-24;**修回日期**:2018-09-18

基金项目:教育部人文社会科学研究青年项目(15YJC860001);青岛市社会科学规划项目(QDSKL1701074);山东省自然基金面上项目(ZR2017MG011);山东省社会科学规划项目(17CHLJ16)资助.

作者简介:孙更新(通讯作者),男,1978年生,博士/副教授. 研究方向:数据挖掘、复杂网络. E-mail:sungengxin@qdu.edu.cn

0　引言

近年来，随着 Web 2.0 等技术的兴起，以微博和虚拟社区（如 Facebook、豆瓣网、Twitter、新浪微博等）为代表的社交网络大量涌现，成为网民获取和传播信息、交流观点的主要平台.社交网络是借由互联网进行沟通，由具备共同的兴趣爱好，发表和讨论感兴趣的话题的用户形成的团体.在社交网络中，用户不再是被动地接收信息，而是主动地提供信息、传播信息，通过这种主动的、强交互的方式，完成信息的共享和传播.

网络舆情是指在网络上传播的、具有一定规模的、大众对某一“焦点”、“热点”问题所表现出的有一定影响力、带有一定倾向性的共同意见或言论.社交网络参与人数多、联系频繁，资源信息的传播和获取也更加快捷，从根本上改变了网络用户的联系范围和网络行为.在当前网络与现实社会相互交织的环境下，网络舆情已经深刻介入现实生活，形成了崭新的网络舆论场，极大地影响着舆情事件的走势.

舆情信息传播一直是传播学的研究热点问题.不同领域的研究人员对舆情信息传播的研究已经积累了很多的方法和结论.Kermark 早在 1927 年就提出了易染-感染-免疫（susceptible-infected-recovered，SIR）模型[1]，该模型最初用于研究流行病的传播方式，后来广泛应用于网络舆情传播的研究.以该模型为基础建立的网络舆情传播模型，对传播主体和传播载体的假设都过于理想化，造成了传播模型与真实情况之间具有较大差距.随着复杂网络理论的兴起，基于复杂网络拓扑结构和动力学性质的舆情传播模型研究产生了很多新的成果.Zanette 利用平均场近似性研究了小世界网络[2-3]和无标度网络[4]上的信息传播模型，发现复杂网络中信息所能影响的人数比例要比随机网络中的小.刘宗华等则进一步的说明了随机网络是最易进行信息传播的网络[5-6].Kesten 等利用概率理论，将人群视为布朗运动的多粒子系统，研究了信息在移动人群中的传播过程[7].汪小帆等研究了具有幂律分布和可变聚集系数的无标度网络上的舆论传播行为，得出了网络的聚集系数越高，越能抑制信息传播的结论[8].胡晓峰等则利用复杂网络模型研究了论坛和博客等传播媒介对于控制舆情传播的作用和影响[9-10].

本文以仓室模型为基础，根据信息传播主体间存在的多种关系，对传播主体间的兴趣匹配进行预测，进而构建基于兴趣匹配的网络舆情传播模型.该模型对网络中具有相似兴趣的用户进行归类，从而确定了不同用户间进行信息传播时兴趣差异对传播造成的影响，而已有社交网络舆情信息传播模型，往往忽略了信息传播的有效目标应该是兴趣匹配的用户这一事实，因而造成了模型与真实情况具有较大差异.本文通过设置影响网络舆情传播的参数，对模型进行仿真分析，探寻网络舆情传播中的规律，进而为网络舆情传播的控制提供理论依据.

1　社交网络用户显式和隐式关系分析

社交网络是一种典型的复杂网络[11-12]，主要研究个体之间的联系以及行为活动.社交网络中，用户间的联系一部分来自于现实世界中的关系，并且通过网络交流得到进一步强化.另一部分来自于网络用户共同的网络行为活动，并逐步形成了网络社团结构，因此社交网络必然是一个多关系网络，作为网络节点的用户之间也必然存在着多种关系.

在社交网络中，用户可以请求添加另一用户为好友.如果双方成为好友，则用户之间就形成了较为稳固的关系，以后在该社交网络中可以自由交流.这种关系通常源于线下原本就存在的好友关系，是线下关系在社交网络中的延伸，也可能是通过在线交流后形成的用户关系的固化.这种用户间存在的交流关系是显式存在的，称之为显式关系.社交网络中用户的行为会体现出用户的兴趣和喜好，如用户可能会对感兴趣的资源进行评论或分享.虽然众多用户的兴趣爱好各有不同，但不同用户的兴趣会有一定的相似性.这些具有相同兴趣的用户间可能不存在显式关系，但由于他们之间存在隐含的相同的兴趣，以后有可能建立显式关系.这种由用户行为表现出的共同兴趣特征的用户隐式的关系，称之为用户隐式关系.例如，在微博系统中，依据微博用户的行为方式，微博用户之间至少存在关注、回复、转发和阅读四种显式关系，进一步对微博内容和微博用户间的互动行为进行分析，可以从中发现用户的兴趣偏好，从而找到用户之间存在的各种隐式关系.

在社交网络的舆情传播过程中，用户隐式关系和用户显式关系所起的作用也不尽相同.显式关系是用户间进行舆情传播的充分条件，即不存在显式

关系的用户之间是不可能直接进行舆情传播的.隐式关系隐含于用户行为之中,不像显式关系那样存在真实的交流关系,隐式关系是虚拟的,是为了研究用户兴趣以及彼此之间舆情传播的可能性而假想存在的,因此社交网络中隐式关系的边是一条“虚边”.由于用户在舆情传播过程中往往只会对自己感兴趣的信息进行传播,所以必须考虑隐式关系对舆情传播的影响.

1.1 数据来源及采集

微博系统按照用户标签、微博文本、关注用户列表等信息推送消息,充分利用这些信息,可以挖掘用户的兴趣偏好,建立用户兴趣匹配模型,从而预测用户间的兴趣匹配关系.本文以新浪微博为研究对象,采用编程及网络爬虫软件获取实验数据.利用新浪微博 API,从种子用户出发,在 2017 年 1 月到 2017 年 2 月间,连续爬取一个月,最后得到 577 467 个用户的属性信息,共获得微博 36 271 229 个.其中用户之间具有相互关注关系的用户数为 145 776 个,这是具有显式关系的用户数目.

1.2 数据处理

采集到的微博用户属性信息数据包括:用户 ID、地区、用户名、关注(关注用户数)、标签、关注列表等字段.用户微博文本信息包括:用户发表和转发的微博,采集到的原始数据如表 1 所示.

表 1 采集的微博原始数据

Tab.1 Raw data collected from micro-blog

微博用户属性信息数据	用户微博文本信息
2549228714,英国那些事儿,男,,,英国,hereinuk,391,821,765,,, http://weibo.com/hereinuk,"英国趣闻,事儿君,英国那些事儿,留学生,趣事,英国 ",2011 年 11 月 21 日,,,,," 6128774301,3149183682,6073974932,6034931374,5124266126,3503381065,5693047492,5081752463,2647236162,6124463229,....", 0,2279,20231	<comment><content>“盗猎者不问罪过一律射杀!”这家以独角犀牛闻名的野生动物保护区,对待偷猎者的态度也是有点狠……</content><time>2017-2-12 19:59</time><repostsCount>758</repostsCount><commentsCount>1022</commentsCount></comment>

为了准确提取用户兴趣,需要对微博数据信息进行提取和分词处理.本文采用中科院 NLPIR 中文分词 Java 版作为数据的分词工具.根据工具提供的中文语料库和分词模型,得到的分词效果如表 2 所示.

表 2 微博原始数据分词结果

Tab.2 Segmentation results of micro-blog raw data

分词前	分词后
英国趣闻,事儿君,英国那些事儿,留学生,趣事,英国,盗猎者不问罪过一律射杀!”这家以独角犀牛闻名的野生动物保护区,对待偷猎者的态度也是有点狠	英国,趣闻,事,君,那些,留学生,趣事,盗猎者,不问,罪过,一律,射杀,这家,独角,犀牛,闻名,野生,动物,保护区,对待,偷猎者,态度,也是,有点,狠

对微博数据信息进行分词处理后,去掉分词结果中的一些无意义词,如“那些”、“不问”、“一律”、“也是”、“有点”等,提取关键词,如“英国”、“留学生”、“盗猎者”、“动物”、“保护区”等.

1.3 数据分析

根据采集得到的关注用户列表,利用复杂网络理论,以用户之间的关注关系为边,以微博用户为节点,构建新浪微博用户显式关系网络.网络中用户节点度的大小体现了用户的活跃程度,其分布情况反映了用户关系的分布.使用 Gephi 工具统计得到网络中节点度为 1 的用户占 8%,节点度小于等于 10 的用户占 24%.其分布如图 1 所示.

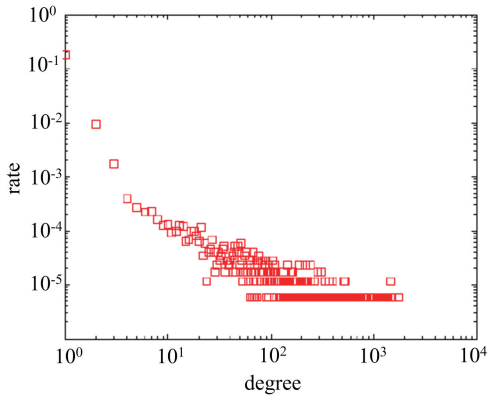

图 1 新浪微博用户显式关系网络节点度分布

Fig.1 Degree distribution of Sina micro-blog user explicit relationship network

参照主流社交网络对用户兴趣的划分标准,可以将用户兴趣分为(美食,教育,娱乐,体育,时尚,财经,科技、文化,军事,读书,汽车,音乐,游戏,星座,影视,购物,摄影,宠物,新闻,搞笑,生活)等 21 类.通过文本分词,提取关键词,并根据这 21 个兴趣类,将每个用户标签内的关键词与每个兴趣类进行对应,其中,每个兴趣类会对应多个关键词.

本文定义用户兴趣向量为:美食,教育,娱乐,…,搞笑,生活.通过每个兴趣类对应的关键词,可以将代表用户兴趣的标签和微博文本信息定义成每个用户的兴趣向量,从而得到用户间共同兴趣标签的数量以及微博文本信息中兴趣类对应的数量.

本文设定,如果两个用户具有 5 个以上共同的兴趣标签,或转发/评论共同的微博数量超过 10 个,或所发微博的文本信息中的兴趣类对应数量超过 10 个,则认定他们之间存在一种基于兴趣匹配的隐式关系.根据用户间兴趣匹配构建的新浪微博用户隐式关系网络的节点度分布如图 2 所示.

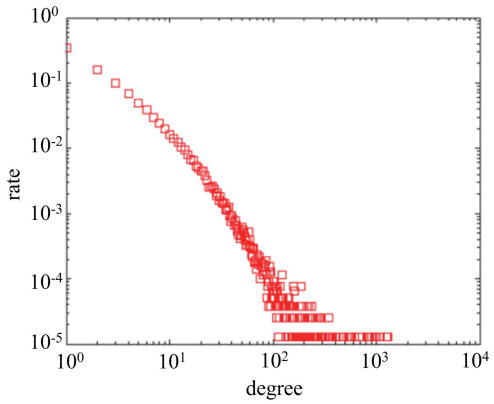

图 2　新浪微博用户隐式关系网络节点度分布

Fig.2　Degree distribution of Sina micro-blog user implicit relationship network

通过图 2 可以看出,网络中节点度为 1 的用户占 35.1%,节点度小于等于 10 的用户占 85.7%.这说明网络中兴趣匹配度很高的用户并不多.

通过上述数据分析结果可以看出,新浪微博用户显式关系网络和隐式关系网络的度分布都服从幂率分布,都呈现无标度特性,这与已有的社交网络研究结果一致[13].

2　用户兴趣匹配预测算法

爬取的微博数据包含每个用户的标签信息和微博文本信息,通过对这些数据信息进行分词和关键词提取,可以得到反映用户兴趣的兴趣类.用户 i 的兴趣向量可以描述为

$$T_i=(T_{Li},T_{Ci}).$$

式中,T_{Li} 表示从用户 i 的用户标签所得到的兴趣向量,T_{Ci} 表示从用户 i 的微博文本信息所得到的兴趣向量.其对应的权值向量为

$$W_i=(W_{Li},W_{Ci}).$$

式中,W_{Li} 代表用户 i 的用户标签兴趣度向量,W_{Ci} 代表用户 i 的微博文本兴趣度向量.

由于用户 i 的标签信息和微博文本信息中包含多个特征值,则 T_{Li} 和 T_{Ci} 可以用如下向量形式表示:

$$T_{Li}=((T_{Li1},W_{Li1}),\cdots,(T_{Lim},W_{Lim})),$$
$$T_{Ci}=((T_{Ci1},W_{Ci1}),\cdots,(T_{Cin},W_{Cin})).$$

此时用户兴趣向量和兴趣度向量分别为

$$T_i=(T_{Li1},T_{Li2},\cdots,T_{Lim},T_{Ci1},T_{Ci2},\cdots,T_{Cin}),$$
$$W_i=(W_{Li1},W_{Li2},\cdots,W_{Lim},W_{Ci1},W_{Ci2},\cdots,W_{Cin}).$$

由上述定义可知,用户的微博文本兴趣是由微博文本信息中的兴趣类对应的特征值表征,用户标签兴趣是由用户标签信息来表征.如果用户的兴趣中不包含某项兴趣类,则在其兴趣度向量中将对应的兴趣度值设置为 0.本文将依据用户兴趣向量和兴趣度向量,计算用户间的兴趣相似性,从而进行用户兴趣匹配预测.

2.1　用户兴趣的度量

假设用户 i 的关注列表中共有 v 个用户,对这 v 个用户的标签求并集,该并集即用户 i 的关注列表标签集.对于用户标签兴趣,本文认为用户 i 对其 m 个兴趣类的偏好程度,可以通过该用户的关注列表中全部用户的标签集 X 内标签出现的频率来表征.以兴趣类在关注列表中出现的频率来度量用户 i 对其标签中的兴趣类 j 的兴趣度,可以表达为

$$W_{Lij}=\frac{\sum c_j}{v} \tag{1}$$

式中,$\sum c_j$ 表示关注列表内包含兴趣类 j 的用户数量,v 表示关注列表内用户的数量.

对于用户的微博文本兴趣,所发微博时间越接近当前时刻越能体现用户的当前兴趣,这种现象类似于人类行为动力学中的兴趣衰减函数[14],因此可以通过定义兴趣衰减函数来描述用户的微博文本兴趣.兴趣衰减函数定义为

$$x(t)=\frac{1}{(1+kt)},t\in(0,\infty) \tag{2}$$

式中,k 表示衰减速率.

衰减函数表示在时间段$[t_0,t]$内记忆量随时间的变化范围.假设用户在 t_0 时刻的兴趣度为 $P(t_0)$,根据兴趣衰减函数,从 t_0 到 t 时刻用户的兴趣度将会降低到

$$P(t)=\frac{P(t_0)}{(1+k_t(t-t_0)} \tag{3}$$

式中，k_t 表示在时间段 t_0 到 t 内的兴趣的衰减速率. 基于衰减函数可以计算得到用户的微博文本兴趣度向量.

2.2 用户相似性计算

根据式(1)和式(3)定义的用户兴趣的度量，可以得到每个用户的用户兴趣向量 T_i 和兴趣度向量 W_i，并利用这两个向量进行用户相似性计算.

本文以余弦相似性度量来计算任意两个用户的兴趣相似性，具体公式为

$$S_{ij}=\cos(T_iW_i,T_jW_j)=\frac{T_iW_i * T_jW_j}{\sqrt{(T_iW_i)^2}+\sqrt{(T_jW_j)^2}} \tag{4}$$

通过式(4)计算用户间的兴趣相似性，S_{ij} 值较大，表示用户之间具有兴趣匹配关系的可能性较大；反之，则可能性较小.

本文将爬取所获得的新浪微博用户数据分为 90%的训练集和 10%的测试集，并根据用户的标签、微博文本内容和关注列表来获取用户的兴趣.为了验证本文所提算法的预测效果，选取经典的用户兴趣匹配预测算法 TF-IDF 和 LDA[15]，语义分析用户兴趣匹配预测算法 DPLSA [16]以及融合网络拓扑和微博内容的用户兴趣匹配预测算法 TFP[17]，与本文的用户兴趣匹配预测算法进行对比分析，并使用 AUC 和 Precision 评价指标[18]对这些算法进行评价，得到的预测效果分别如图 3 和图 4 所示.

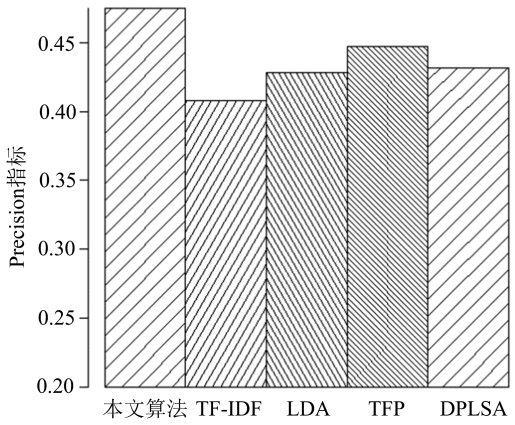

图 3 Precision 指标下不同用户兴趣匹配预测算法的评价值对比

Fig.3 Comparison of evaluation values of different user interest matching prediction algorithms under Precision

比较这些算法的评价值可以发现，本文所提用户兴趣匹配预测算法的预测效果最好，能够较为准确地描述用户间的兴趣匹配关系，并提高用户间兴趣匹配关系预测的准确性.

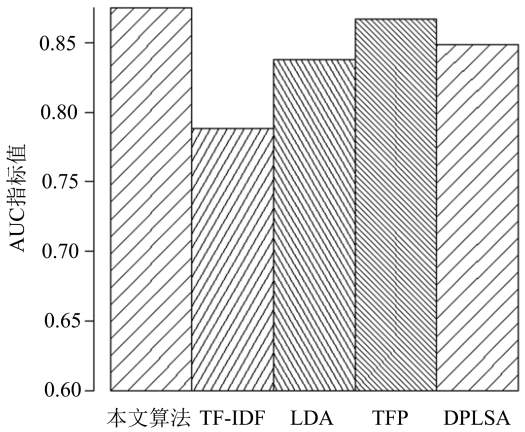

图 4 AUC 指标下不同用户兴趣匹配预测算法的评价值对比

Fig.4 Comparison of evaluation values of different user interest matching prediction algorithms under AUC

3 基于兴趣匹配的网络舆情传播模型

3.1 仓室模型

运用传染病动力学模型进行网络舆情传播的研究是应用最为广泛的方法，其中应用最多的是“仓室”模型[19].“仓室”模型是一类用于模拟传染病传播过程的数学模型的统称，SIR 模型是其中最经典的代表，该模型可以模拟传染病的传播过程，进而预测传染病的爆发及传播规律.

SIR 模型把传染病流行范围内的人群分成 3 类：S 类，易感者(susceptible)，指未得病者，但缺乏免疫能力，与感病者接触后容易受到感染；I 类，感病者(infective)，指染上传染病的人，它可以传播给 S 类成员；R 类，移出者(removal)，指被隔离，或因病愈而具有免疫力的人.SIR 模型的演化规则可以描述如下：

(Ⅰ)易感者以传染率 λ 被某个感病者所感染，并从易感状态转变成为感病状态；

(Ⅱ)处于感病状态的个体以免疫率 γ 被治愈，并获得免疫能力；

(Ⅲ)设 $S(t)$，$I(t)$，$R(t)$ 分别为 t 时刻易感者、感病者以及移出者在种群总人口中所占的比例，即 3 类不同状态个体的密度，则 $S(t)+I(t)+R(t)=1$.

该模型的传播过程可以用以下微分方程组描述：

$$\left.\begin{aligned}\frac{\mathrm{d}S(t)}{\mathrm{d}t}&=-\lambda I(t)S(t)\\ \frac{\mathrm{d}I(t)}{\mathrm{d}t}&=-\lambda I(t)S(t)-\gamma I(t)\\ \frac{\mathrm{d}R(t)}{\mathrm{d}t}&=\gamma I(t)\end{aligned}\right\}\tag{5}$$

3.2　基于兴趣匹配的网络舆情传播模型

本文提出的基于兴趣匹配的网络舆情传播模型是在 SIR 模型的基础上，考虑用户间兴趣匹配关系，定义各状态之间的转移过程和转移概率.

在该网络舆情传播模型中，将网络中的全部用户划分为以下 4 种状态：易感状态 S，接收状态 E，传播状态 I 和免疫状态 R.易感状态是指用户从未获知目标舆情信息，即对该舆情信息处于完全未知；接收状态表示用户已经通过其他用户的转发而获知了目标舆情信息，但还未转发该舆情信息；传播状态是指用户已将目标舆情信息转发；免疫状态是指用户完全对目标舆情信息失去兴趣，并且永远不会对其进行转发.

根据网络中用户间兴趣匹配关系以及网络中节点度的分布，上述 4 类节点的状态转移过程定义如下：

(Ⅰ)设 $N(k,t)$为 t 时刻网络中度为 k 的节点总数，$S(k,t)$，$E(k,t)$，$I(k,t)$及 $R(k,t)$ 分别表示 t 时刻网络中度为 k 的易感节点、接收节点、传播节点及免疫节点的密度，即上述 4 类节点的数量分别在 $N(k,t)$中所占的比例，且 $S(k,t)+E(k,t)+I(k,t)+R(k,t)=1$；

(Ⅱ)当易感节点 S 接触到一个传播节点 I，则该易感节点以概率 p_{se} 转变为接收节点 E，p_{se} 称为感染概率；

(Ⅲ)根据接收节点 E 与传播节点 I 之间兴趣匹配关系，接收节点 E 以概率 p_{ei} 转变为传播节点 I，则 p_{ei} 称为接收节点 E 对目标舆情信息的转发概率；

(Ⅳ)根据接收节点 E 与传播节点 I 之间兴趣匹配关系，接收节点 E 以概率 p_{er} 转变为免疫节点 R，则 p_{er} 称为接收节点 E 对目标舆情信息的直接免疫概率；

(Ⅴ)传播节点 I 以概率 p_{ir} 转变为免疫节点 R，则 p_{ir} 称为传播节点 I 对目标舆情信息的免疫概率；

(Ⅵ)免疫状态为网络中的吸收状态，即进入免疫状态的节点，其状态不再发生改变.

根据上述状态转换规则，社交网络中基于兴趣匹配的网络舆情传播模型如下：

$$\frac{\mathrm{d}S(k,t)}{\mathrm{d}t}=-p_{se}k\theta(t)S(k,t),$$

$$\frac{\mathrm{d}E(k,t)}{\mathrm{d}t}=p_{se}k\theta(t)S(k,t)-p_{ei}E(k,t)-p_{er}E(k,t),$$

$$\frac{\mathrm{d}I(k,t)}{\mathrm{d}t}=p_{ei}E(k,t)-p_{ir}I(k,t),$$

$$\frac{\mathrm{d}R(k,t)}{\mathrm{d}t}=p_{er}E(k,t)+p_{ir}I(k,t).$$

式中，$\theta(t)$ 表示 t 时刻网络中任意一条随机边与传播个体相连接的概率.设 $P(k)$ 为网络的度分布函数，$<k>$为网络的节点平均度，则

$$\theta(t)=\frac{\sum_{k}kP(k)I(k,t)}{\langle k\rangle}\tag{6}$$

3.3　模型仿真与分析

本文使用 Matlab 作为工具对构建的网络舆情传播模型进行仿真，在仿真过程中经过 200 次传播迭代后，舆情信息在网络中的传播过程基本趋于稳定，通过对仿真结果进行分析，可以得到以下结论.

3.3.1　不同类型节点随时间的变化趋势

$S(t)$，$E(t)$，$I(t)$ 及 $R(t)$ 分别表示 t 时刻网络中 4 类节点的密度.设置模型参数 $p_{ei}=0.2$，$p_{er}=0.1$，$p_{ir}=0.3$，$p_{se}=1$，并选择网络中度最大的节点作为舆情信息传播的初始节点，此时，网络中不同类型节点随时间的变化趋势如图 5 所示.

从图 5 可以看出，网络中的易感节点数量在舆情传播初期快速减少，这是因为一旦某个节点处于传播状态，网络中所有与该节点有相连边的其他节点都将转变为接收状态，这正体现了社交网络中舆情信息的“裂变式”传播模式.网络中的接收节点数量在舆情传播初期会快速增长，并且在极短时间内达到最大值，之后随着易感节点数量的减少以及接收态节点向传播态和免疫态的转变，其数量会随时间逐渐减少，并最终趋近于 0.网络中的传播节点与接收节点的变化趋势类似，也是在传播初始阶段快速增加并达到最大值，之后将逐渐减少并最终趋近于 0，只是变化的各阶段在时间上滞后于接收节点.网络中的免疫节点在传播初始阶段会逐渐增加，最终将趋近于 1，即网络中的所有节点最终都将转变为免疫状态，这也反映出免疫状态将成为网络的吸收状态.

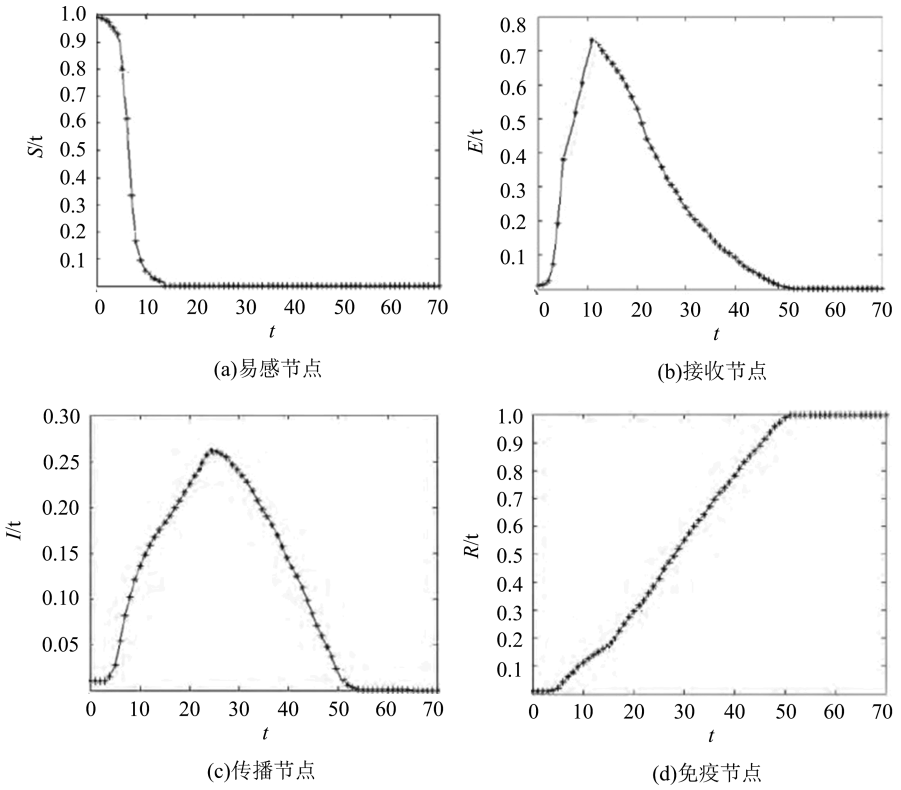

图 5 不同类型节点随时间的变化关系

Fig.5 The relationship between different types of nodes over time

3.3.2 用户间兴趣匹配关系对传播过程的影响

用户节点之间兴趣匹配关系主要影响转发概率 p_{ei} 和直接免疫概率 p_{er}.图 6 显示了转发概率 p_{ei} 取不同值时,传播节点数量和免疫节点数量随时间的变化趋势情况.

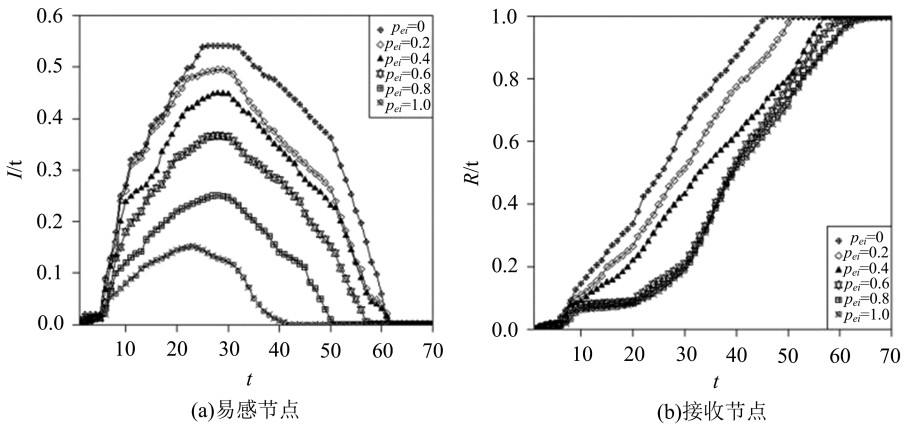

图 6 转发概率对传播节点及免疫节点数量的影响

Fig.6 The influence of forwarding probability on the number of propagation node and immune node

从图 6 的曲线变化可以看出,在网络达到稳态之前,p_{ei} 取值越大,则 $I(t)$ 的值也会越大,而 $R(t)$ 的值会越小.这是因为 p_{ei} 表示接收节点转变为传播节点的概率,p_{ei} 值的增大,表明处于接收状态的节点,由于与传播节点间的兴趣匹配一致,而导致转发舆情信息的概率的增加.此外,$I(t)$ 的值趋近于 0 的时间,会随着 p_{ei} 的值的增大而增长,这是因为随着 p_{ei} 的值的增大,网络中传播节点的数量也将随之增多,这会导致需要更长的时间才能使传播过程达到最终的稳定状态.

图 7 显示了直接免疫概率 p_{er} 的取值对网络中传播节点数量以及免疫节点数量随时间变化的影响.