doi: 10.12011/1000-6788(2018)08-2135-12 中图法分类号: TP391.1 文献标志码: A

融入网络结构与社交习惯的不对称用户关系强度计算

琚春华 [1,2], 陈 彦 [2], 鲍福光 [1,2]

(1. 浙江工商大学 现代商贸研究中心, 杭州 310018; 2. 浙江工商大学 管理工程与电子商务学院, 杭州 310018)

摘 要 伴随着大量用户内容的创建和交换, 社交网络平台中产生了大规模的互动数据和复杂的用户关系, 受到了越来越多研究者的关注. 但是现有对关系强度研究多是从用户特征属性相似度和社交行为两方面进行, 忽略了网络结构对关系强度的影响, 且并未考虑社交行为存在的方向性和习惯性问题. 基于此, 本文提出了不对称的社交网络用户关系强度计算方法 (DSTS-ATI), 该方法融合用户特征属性相似度、网络结构连接强度、社交行为交互强度三个维度来综合计算用户关系. 在计算网络拓扑结构连接强度时, 不仅考虑了用户间邻居节点数, 还考虑了邻居节点连接边数. 社交互动行为发生的方向性和习惯性, 会影响用户对关系强度的感知, 因此本文计算出不同社交行为的贡献权重, 从交互双方感知用户社交强度. 实验证明, 本文提出的不对称用户关系强度的方法能够提高用户关系强度预测的准确性, 有助于微博意见领袖的发现和信息传播机制的研究.

关键词 不对称性; 网络结构; 社交习惯; 用户关系强度

A dissymmetrical tie strength estimation method based on topology of networks and habituation of social interactions

JU Chunhua[1,2], CHEN Yan[2], BAO Fuguang[1,2]

(1. Contemporary Business and Trade Research Center, Zhejiang Gongshang University, Hangzhou 310018, China;
2. School of Management and E-business, Zhejiang Gongshang University, Hangzhou 310018, China)

Abstract As the product of web2.0 era, the emergence of social network sites make people bridge the traditional social gap caused by the temporal and spatial distance. Its wide spread and rapid development have changed our lifestyles dramatically. With the mass creation and frequent exchange of user generated content (UGC), an amount of interactive data and complex user relationship have been generated in the social network sites. This phenomenon attracts the attention of enormous researchers. However, extant researches of the tie strength mainly focused on the user attributes and social interactions. But they ignored the influence of network structure, and did not take into account the direction and habituation of social interactions. Hence, we propose a dissymmetrical tie strength estimation method based on user attributes, topology of networks and social interactions (DSTS-ATI). From the perspective of topology of networks, we consider not only the number of common neighbor nodes, but also the links between the nodes. The direction and habituation of social interactions affect the users' perception of tie strength. Therefore, this paper gives out the weights of different social interactions bidirectionally. The experimental results show that the proposed method enhances the accuracy of tie strength prediction. Further, it is beneficial to the research of opinion leaders discovery and the information dissemination mechanism.

收稿日期: 2017-03-30

作者简介: 琚春华 (1962–), 男, 汉, 博士, 教授, 博士生导师, 研究方向: 智能信息处理与数据挖掘, 商务智能, E-mail: juchunhua@hotmail.com; 陈彦 (1992–), 女, 汉, 硕士研究生, 研究方向: 智能信息处理与数据挖掘, 电子商务, E-mail: cymail92@163.com; 通信作者: 鲍福光 (1986–), 男, 汉, 博士, 讲师, 研究方向: 智能信息处理与数据挖掘, 电子商务, E-mail: baofuguang@126.com.

基金项目: 国家自然科学基金 (71571162); 国家科技支撑计划项目 (2014BAH24F06); 浙江省哲学社会科学规划课题 (16NDJC188YB)

Foundation item: National Natural Science Foundation of China (71571162); National Key Technology R&D Program of China (2014BAH24F06); Zhejiang Province Philosophy Social Sciences Planning Project (16NDJC188YB)

中文引用格式: 琚春华, 陈彦, 鲍福光. 融入网络结构与社交习惯的不对称用户关系强度计算 [J]. 系统工程理论与实践, 2018, 38(8): 2135–2146.

英文引用格式: Ju C H, Chen Y, Bao F G. A dissymmetrical tie strength estimation method based on topology of networks and habituation of social interactions[J]. Systems Engineering — Theory & Practice, 2018, 38(8): 2135–2146.

Keywords dissymmetry; topology of networks; habituation of social interactions; tie strength

1 引言

随着 Web2.0 的发展, 在线社交网络如 Facebook、Twitter、微博、微信等, 允许用户以分享文字、图片、音频、视频等方式进行信息的交流和传播, 颠覆了我们传统的社交生活. 作为 “互联网 +” 时代最热门的应用, 社交网络因其便捷的接入方法和灵活的参与方式, 成为个人情感抒发、态度观点表达、信息分享互动的主要社交平台, 且受到了越来越多网络用户的青睐. 以社交关系管理为基础, 它的服务范围已经扩展到新闻传媒、应用集成、商务交易等领域 [1], 已然渗入到生活的方方面面.

传统社交网络, 例如人人网、腾讯 QQ, 用户关系的建立是无方向性的, 用户关系的建立是需要相互认证的. 与大多数社交网络在结构上存在很大的不同, 微博用户关系的形成是用户依据个人兴趣来选择关注对象, 以此来获取信息资讯, 但此时形成的关系多为单方向的关注关系. 与此同时, 用户关系的形成还依赖于线下好友、同学、同事、家人等亲近的线下熟人关系, 现实世界存在的好友关系很大程度上迁移到社交网络中, 通常表现为在微博中相互关注 [2]. 此外, 微博作为开放性的社交平台聚集了大量的用户, 这些用户在文化背景、社交目的上都不尽相同, 并且以一对多的方式进行互动交流 [3]. 如何挖掘这些隐藏在用户大规模数据和复杂关系背后的社会经济价值, 已经成为学者们当下研究的热点.

基于便捷的操作和多样化的功能服务, 微博已经成为国内最热门的社交应用之一. 据《2016 年度微博用户发展报告》显示, 截至 2016 年 9 月, 微博月活跃用户数 (MAU) 已经达到 2.97 亿人, 日均活跃用户数 (DAU) 达到 1.2 亿. 除转发、评论、点赞、收藏等互动体验外, 微博平台还为用户提供了更具多样化和趣味性的服务功能. 例如, 长微博、打赏功能的开发以及视频微博的推广, 也进一步使微博用户之间的交互更加多元化. 微博支付的持续推广, 直播、购物、充值、粉丝头条、打赏、付费阅读、微博会员以及送花等颇具特色业务的发展, 使微博的使用价值和商业价值进一步提升. 因此, 准确衡量社交网络用户关系强度对推进网络应用发展和个性化服务有着重要意义.

2 文献综述

关系强度的概念是由 Granovetter 于 1974 年首先提出, 他从交互时间、情感强度、亲密程度和互惠性四个维度对其进行了定义, 并把关系强度分成强关系和弱关系, 创立了弱关系理论 [4]. 此后, 研究者们又进一步发现, 弱关系有利于信息的传播 [4] 和求职 [5]; 且在推荐经验因素影响下, 弱关系使推荐结果更具有说服力 [6]. 而另一方面, 社会交换理论表明, 来源于强关系的信息有更高的经济价值 [7]; Wellman 和 Wortley 发现强关系能够提供更多情感支持, 例如提供家庭关系问题的处理意见 [8]; 但在组织变革过程中强关系会带来更大的阻力 [9].

随着社交网络的兴起和广泛应用, 越来越多的研究者将注意力转移到在线社交网络中的用户关系, 取得了丰富的研究成果, 他们发现社交网络用户关系对链路预测 [10]、好友推荐 [11]、消费购买 [12,13]、信息传播 [14] 扮演着重要角色. 现有对在线社交网络用户关系强度的研究主要分为四个方面: 1) 关系强度影响因素研究; 2) 关系类别研究; 3) 关系强度预测方法研究; 4) 关系强度应用研究. Kahanda 等人从用户的属性特征 (性别、婚姻状况、社会地位等), 事物特征 (群成员上传照片等), 互动特征 (分享行为) 来判断好友关系的亲密度, 并且发现互动特征是关系强度预测中最重要的特征 [15]. 文献 [16] 通过计算用户属性相似度和互动强度, 对用户关系进行了强弱关系的划分. 文献 [17] 在研究用户关系时引入了社会网络理论, 建立用户评级模型用于评价用户推荐能力. 文献 [18] 提出了微博动态相似度计算和动态交互相关性计算方法.

在对社交网络用户关系强度预测的研究中, 用户个人属性特征是常见的影响因素. Xiong 等人综合考虑了用户基本属性相似度、姓名共现性和交互频率后, 提出了图模型方法来预测关系强度 [19]. 文献 [20,21] 基于用户个人信息和交互行为两个因素对用户关系强度进行了研究. 但是, 这些方法存在一定的缺陷, 用户隐私保护意识强, 用户个人信息填写不完整, 研究者面临属性稀疏性等问题, 实验过程受到较大约束. 文献 [22] 认为微博短文本能在一定程度上表明用户的兴趣及属性. 因此, 本文在计算用户相似度时综合考虑了用户个人信息和微博短文本信息.

用户与关注对象的亲疏程度与现实世界中的人物关系具有极强的同构性, 这就为社交网络中用户关系的研究提供了思路. 文献 [23] 通过用户共同好友数等因素来测量关系强度, 提出了融入社交关系强度的用户地理位置预测方法. 文献 [24] 认为社交网络结构 (网络拓扑结构、社交圈) 是用户关系强度重要影响因素, 节点间连接紧密性能体现出用户关系强度的差异. Chulyadyo 等人利用共同相邻节点数来计算节点间的连接强弱 [25]. 但是, 从用户共同邻居节点 (CN) 来衡量网络结构连接强度的信息量是不够的, Ravasi 等人发现如果两个节点间有共同邻居节点, 且共同邻居节点间也存在连接边, 那么这两个节点之间存在连接边可能性更大 [26]. 基于此, 本文在研究网络结构连接强度时, 不仅考虑了邻居节点, 还考虑了邻居节点是否存在连接边的情况.

用户在微博上的社交行为表明用户之间存在着紧密的联系, 因此用户社交行为是用户关系强度研究不可忽略因素. 但是现有计算社交行为对关系强度影响的研究都不够深入, 例如只关注交互活动发生与否 [19], 或者只考虑了交互行为发生的总次数 [27,28], 并未对社交行为的方向性和习惯性进行区分. 例如, 用户 A 是一个与其好友交互十分活跃的用户, 习惯性点赞, 那么用户 A 与用户 B 所产生的高频率互惠行为不能够充分体现出用户 A、B 之间存在强关系. 且用户之间的关注关系有双向关注和单向关注两种, 交互行为是存在明显方向性的. 社交行为发生的方向不同, 以及在这个方向上社交行为的次数、时间的差异使得用户感知出不同的关系强度, 因此, 对于用户关系判断需要综合考虑社交互动行为发生的方向性和习惯性. 本文在综合考虑用户互动的方向性和社交行为习惯的因素后, 提出了基于社交行为习惯的不对称交互强度计算方法.

3 融合特征属性、网络结构与社交习惯的关系强度计算模型

3.1 问题描述

为了能够对本文的研究任务进行有效说明, 先对相关概念进行定义.

定义 本文把微博增广社交网络定义成一个元组的集合, $G=\{V,E,UA,UI,RS\}$. 其中, V 是指所有用户集, E 是指用户之间通过关注形成的连接边集, UA 为用户特征属性, UI 为用户社交互动行为集, RS 为用户关系强度集, $RS=\left\{RS^l,RS^u\right\}$, RS^l 为用户标注的关系强度集, RS^u 为未标注的关系强度集. 用户特征属性 $UA=\{UB,UT\}$, 其中 UB 为用户背景属性集, 定义为 $UB=\{u\_gen,u\_loc,u\_occ,u\_edu,\ u\_tag\}$, u_gen 指代性别、u_loc 指代所在地、u_occ 是职业信息、u_edu 是教育信息、u_tag 为标签信息. UT 是微博文本属性是指用户发布的短文本内容特征词向量, 定义为 $UT=\{w_0,w_1,w_2,\cdots,w_n\}$, n 是特征词个数. 对于社交互动行为的研究, 考虑用户之间的点赞、转发、评论、@ 提及行为四种情况, 因此把用户社交互动行为定义成 $UI=\{u\_b_1,u\_b_2,u\_b_3,u\_b_4\}$.

基于以上定义, 本文研究的问题可以进行实例化说明, 即给定一个增广社交网络, 预测出未知的用户关系强度, 见公式 (1).

$$f:G=\left(V,E,UA,UI,RS^l\right)\rightarrow RS^u. \tag{1}$$

3.2 用户特征属性相似度计算

现有研究表明职业背景、教育经历相似且相互关注的用户在现实生活中很可能是同事或者同学关系, 交互行为更加频繁 [29]; 社交网络用户更喜欢与地理位置相近或者兴趣偏好相似的人交流 [30]. 为此, 本文爬取微博用户的个人背景资料信息, 包括性别、所在地、职业、教育经历、标签, 定义为 $UB\,(u)=\{u\_gen,u\_loc,u\_occ,u\_edu,u\_tag\}$, 利用 Jaccard 公式计算用户个人背景资料相似性 $sim_{ub}\,(u,v)$, 见公式 (2).

$$sim_{ub}\,(u,v)=\frac{UB\,(u)\cap UB\,(v)}{UB\,(u)\cup UB\,(v)}. \tag{2}$$

但出于隐私保护心理, 用户在注册时不愿透露过多个人信息, 个人资料填写完成度较低 [21], 所以度量用户特征属性相似度还需要考量微博文本信息. 本文对用户 U 微博短文本内容进行了分词和去词操作生成语料库, 提取相应文本的特征词, 利用 TFIDF 公式计算出特征词权重后, 筛选出最终的微博短文本特征词, 定义为 $UT\,(u)=\{w_0,w_1,w_2,\cdots,w_n\}$. 采用经典余弦公式计算用户微博短文本相似度 $sim_{up}\,(u,v)$, 见公式 (3).

$$sim_{up}\,(u,v)=\frac{UT\,(u)\cdot UT\,(v)}{\|UT\,(u)\|\times\|UT\,(v)\|}. \tag{3}$$

综上所述, 用户的相似度 $SIM\,(u,v)$ 计算见公式 (4).

$$SIM\,(u,v)=\lambda_1\times sim_{ub}\,(u,v)+\lambda_2\times sim_{up}\,(u,v)\,. \tag{4}$$

其中, $\lambda_1+\lambda_2=1$, $0\leq\lambda_1\leq 1$, $0\leq\lambda_2\leq 1$. 对于参数 λ_1、λ_2 的取值, 本文以 0.1 为单位长度, 计算 λ_1、λ_2 取不同值对用户相似度的影响, 具体取值情况详见 4.4.4 节内容.

3.3 网络结构连接强度分析

微博用户通过 “关注” 行为与其他用户建立联系, 构建了复杂的社交网络. 作为小世界特性明显的复杂网络 [31,32], 微博中存在许多错综复杂的社群结构. 这些社群结构体现出用户连接关系的本质 - 趋同和聚类. 由于微博是一个存在方向性的复杂网络, 微博中的共同邻居节点存在 4 种情况, 如图 1 所示, 图 1-A 中共同邻居节点是种子用户和交互对象都关注的对象, 图 1-B 中共同邻居节点仅关注种子用户, 图 1-C 中共同邻居节点仅关注交互对象, 图 1-D 中共同邻居节点同时关注了种子用户和交互对象. 节点间的共同邻居节点的数目多, 用户间连接越紧密, 用户之间的亲密程度越高 [33]. 此外, 若共同邻居节点间存在连接边, 如图 2 所示, 那么用户隶属于同一个社群的可能性就越大, 建立紧密联系的可行性更大. 基于此, 本文通过共同邻居节点和邻居节点连接边数来衡量节点间连接紧密程度. 为了方便计算说明, 本文将节点间连接的紧密性定义为网络结构连接强度 (LS), 计算如公式 (5)、(6) 所示.

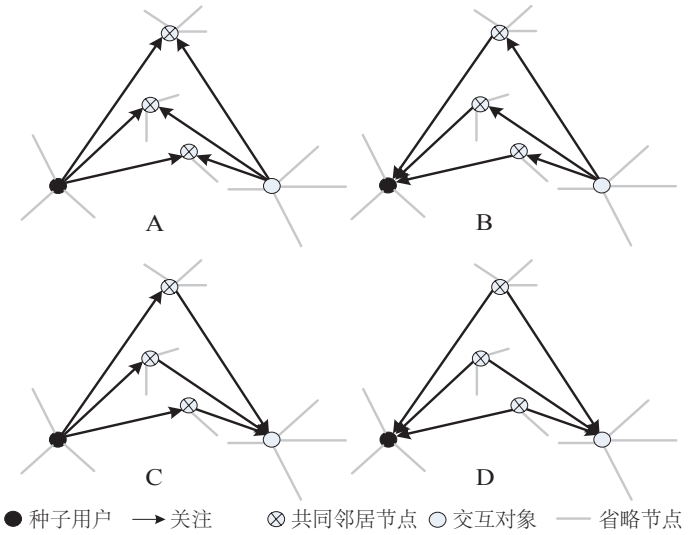

图 1 用户共同邻居节点示意图

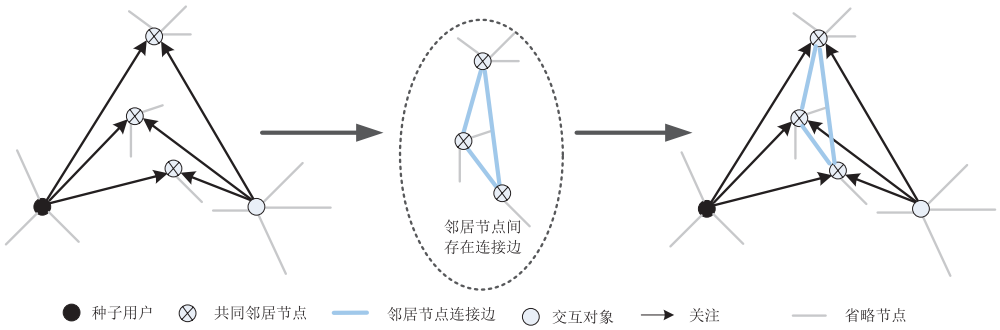

图 2 用户共同邻居节点连接边示意图

$$cnp\,(u,v)=\frac{|\varphi\,(u)\cap\varphi\,(v)|+|\omega\,(u)\cap\omega\,(v)|-|\omega\,(u)\cap\omega\,(v)\cap\varphi\,(u)\cap\varphi\,(v)|}{|\omega\,(u)\cup\omega\,(v)\cup\varphi\,(u)\cup\varphi\,(v)|}. \tag{5}$$

其中, $cnp\,(u,v)$ 表示用户间的共同邻居节点占邻居节点的比例, $\varphi\,(u)$ 表示用户 U 的关注列表, $\varphi\,(v)$ 表示用户 V 的关注列表, $\omega\,(v)$ 表示用户 V 的粉丝列表, $\omega\,(u)$ 表示用户 U 的粉丝列表.

$$LS\,(u,v)=cnp\,(u,v)\times cnl\,(u,v)\,. \tag{6}$$

其中 $LS\,(u,v)$ 为用户间的网络结构连接强度, $cnl\,(u,v)$ 为共同邻居节点间的连接边数.

3.4 基于社交行为习惯的不对称交互强度计算

用户间关注关系的不同, 使得用户社交行为的发生在方向上存在主动和被动差异, 这势必会影响用户之间的关系强度. 因此, 仅从某一方用户的角度去衡量这种关系强度的做法是比较片面的, 忽略了用户社交行为的方向性; 且无论交互活动是单方向还是双方向的, 都会进一步强化用户关系 [33]. 因此, 本文在计算用户交互强度时提出了一种有向的用户交互强度计算方法, 从关系双方分别计算交互强度的感知程度. 此外, 为了确认交互行为对用户关系强度的影响程度, 在计算用户交互强度过程中, 不仅考虑社交行为发生频率, 同时还考虑了不同社交行为习惯对交互关系形成的重要性. 计算出用户该行为的贡献权重, 来综合性考量交互行为对用户感知关系强度的影响大小. 如果该行为常见, 那么就说明这种行为的高频率对高交互强度体现性小, 则降低该行为的权重; 如果该行为不常见, 那么这种行为的高频率对高交互强度体现性大, 就说明该行为对用户关系形成贡献越大.

首先, 先从用户 U 的角度, 计算用户 V 对其交互对象 U 的某种社交行为 b_i 占 U 收到所有该行为的频次比例 BF, 方向为 $u \leftarrow v$, 计算如公式 (7) 所示.

$$BF_{u\leftarrow v,b_i} = \frac{n_{b_i}}{N_{u,b_i}}. \tag{7}$$

其中, b_1 为点赞, b_2 为评论, b_3 转发, b_4 为 @ 提及. n_{b_i} 是用户 V 对 U 主动发生行为 b_i 的次数, 而 N_{u,b_i} 则是 U 收到该种社交行为发生的总次数.

其次, 计算出用户 V 主动发生行为 b_i 的逆交互对象频率 IPF_{v,b_i}, 衡量用户交互行为发生是否集中, 如公式 (8) 所示. 其中, $|p|$ 是指用户 V 所有主动发生的社交行为对象的总人数, $|j : b_i \in p_j|$ 是用户 V 主动发生行为 b_i 的对象人数.

$$IPF_{v,b_i} = \log\left(\frac{|p|}{|j : b_i \in p_j| + 1}\right). \tag{8}$$

最后, 计算出某行为 b_i 对应的贡献权重 $BF \cdot IPF_{u\leftarrow v,b_i}$, 表示该行为对交互活动的重要程度.

$$BF \cdot IPF_{u\leftarrow v,b_i} = BF_{u\leftarrow v,b_i} \times IPF_{v,b_i}. \tag{9}$$

综上所述, 若用 $N_{u\leftarrow v,b_i}$ 表示用户 V 主动对用户 U 发生某行为 b_i 的次数, 则用户 U 感知的与用户 V 之间社交行为强度 $BS_{u\leftarrow v}(u,v)$ 为

$$BS_{u\leftarrow v}(u,v) = \sum_{i=1}^{n} BF \cdot IPF_{u\leftarrow v,b_i} \times N_{u\leftarrow v,b_i}. \tag{10}$$

同理, 也可以计算出用户 V 感知的与用户 U 之间社交行为强度 $BS_{v\leftarrow u}(v,u)$, 具体为:

$$BS_{v\leftarrow u}(v,u) = \sum_{i=1}^{n} BF \cdot IPF_{v\leftarrow u,b_i} \times N_{v\leftarrow u,b_i}. \tag{11}$$

3.5 用户关系强度计算模型

在充分考虑用户个人背景和微博文本相似度、网络结构连接强度、社交行为交互强度对关系强度的影响后, 本文将这三个维度值进行融合, 提出了用户关系强度计算模型, 如图 3 所示. Granovetter 早在 1974 就提出关系强度与交互时间、情感强度、亲密程度和互惠性等有关 [4], 而之后又有大量研文献 [15, 16] 采用了线性组合的方法探究了社交行为、社会距离等对关系强度的影响. 因此, 本文采用线性模型对关系强度进行计算, 是有理论和实践依据的, 具体公式如 (12) 所示.

$$TS = \begin{cases} \alpha \times SIM(u,v) + \beta \times LS(u,v) + \gamma \times BS_{u\leftarrow v}(u,v), \\ \alpha \times SIM(u,v) + \beta \times LS(u,v) + \gamma \times BS_{v\leftarrow u}(v,u). \end{cases} \tag{12}$$

其中, $\alpha + \beta + \gamma = 1$, $0 \le \alpha \le 1$, $0 \le \beta \le 1$, $0 \le \gamma \le 1$.

4 实验

本文以新浪微博为平台, 对微博用户关系强度计算模型进行实验. 由于评价模型需要对用户之间的关系强度进行人工标注, 工作量较大. 且不同领域的微博用户关系强度各不相同, 很难客观对关系强度值进行精确衡量. 因此为了简化实验操作, 本文挑选若干微博活跃用户作为本实验种子用户, 并从这些用户的粉丝列表和关注列表中分别随机抽取若干交互对象, 计算该局部范围内种子用户与其好友的关系强度. 通过让社交

用户自定义关系强度的方式, 形成标准 $Top\text{-}N$ 用户关系强度评分列表. 最终衡量实验结果和标准 $Top\text{-}N$ 用户关系强度列表的差异来评价模型的合理性. 本实验流程如图 4 所示.

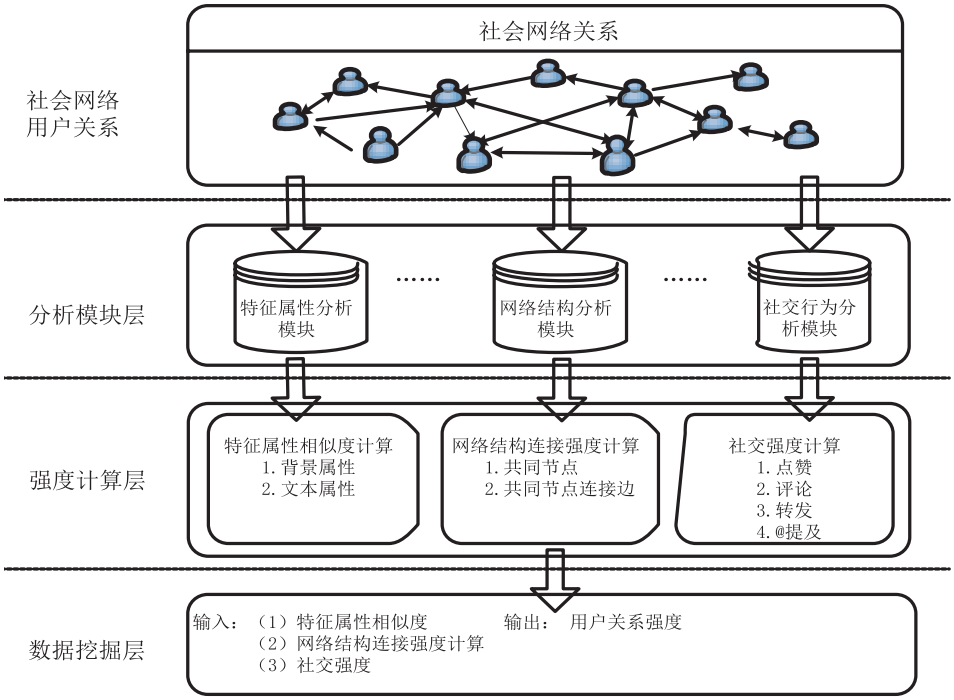

图 3 融人网络结构与社交习惯的不对称用户关系强度计算模型图

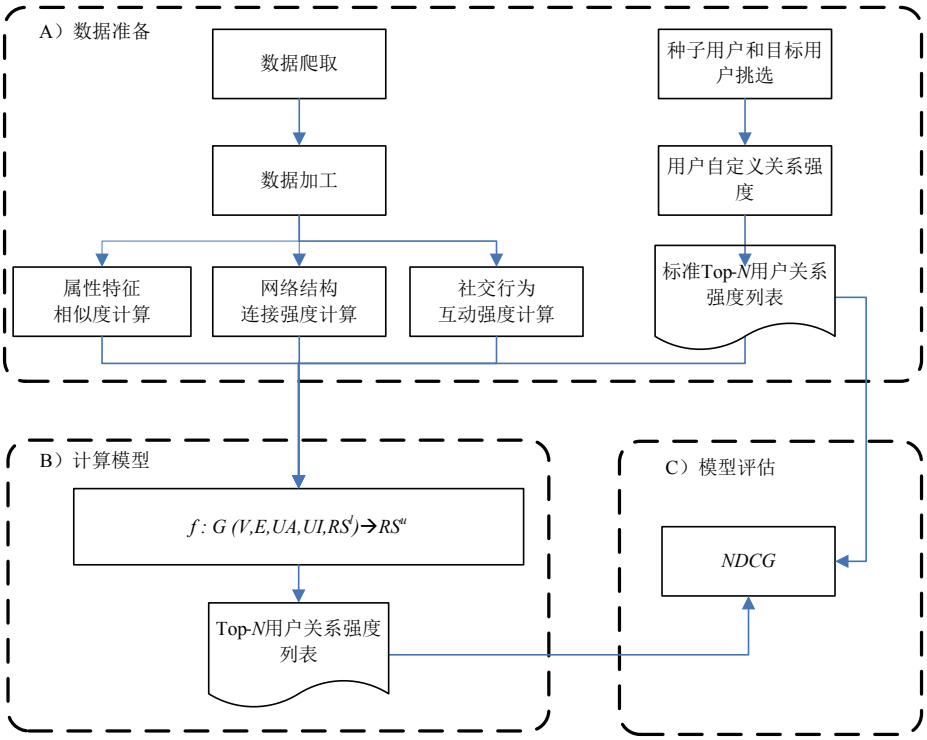

图 4 模型流程图

4.1 数据准备

本实验从体育、美食、电商、旅游、娱乐选取了 20 名微博种子用户 (其中包括 12 女性 (60%) 和 8 男性 (40%), 年龄在 20 岁 ~ 30 岁之间不等, 定期使用微博, 且已至少使用微博一年), 以及从这些种子用户的粉丝列表和关注列表中随机抽取 20 名交互对象, 通过基于模拟登录方式采集了 2016 年 1 月至 6 月期间这 400

个用户的个人背景资料 (性别、所在地、职业、教育、标签)、关注信息 (粉丝列表、关注列表)、微博短文本信息, 并从微博具体内容信息中提取用户间的评论、点赞、转发、@ 提及信息, 总共 13654 文件.

4.2 评价标准

为了对本文建立的多元模型预测结果进行有效评价, 本文将预测结果和人工标注结果进行对比, 选取了 NDCG 指标来评价该模型对用户关系强度排序结果的预测准确性. NDCG (normalized discounted cumulative gain) 归一化折损累积增益, 是信息检索中对检索结果排名常用的评价指标. 该指标在考虑了检索结果重要性的同时, 还充分考虑了排序的相对位置, 增加了评价的有效性, 符合本文对评价排序准确性的诉求, 见公式 (13)、(14).

$$DCG_p = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2 i}. \tag{13}$$

其中, DCG_p 为折损累积增益, rel_i 为种子用户和交互对象之间的关系强度评分, p 为该种子用户的交互对象个数.

$$NDCG_p = \frac{DCG_p}{IDCG_p}. \tag{14}$$

其中, $NDCG_p$ 为归一化折损累积增益, $IDCG_p$ 为理想情况下, 这里即指人工标注结果排名情况下的折损累积增益.

4.3 对比实验

为了进一步体现本文模型对用户关系强度的预测精确性, 本文设计了一系列对比实验. 本文提出了不对称的社交网络用户关系强度计算方法 (DSTS-ATI), 该方法融合用户特征属性相似度、网络结构连接强度、社交行为强度三个维度来综合计算用户关系. 为了体现出本文模型将三个维度加权融合的效果, 将本文的实验结果和文献 [29,30]、文献 [20]、文献 [34] 中所提的用户关系强度计算方法进行了比较. 其中, 文献 [29,30] 提出的方法在计算用户关系强度时, 仅考虑用户背景属性和文本属性相似度 (TS-BP). 文献 [34] 以用户社交行为来衡量用户关系强度 (TS-I); 文献 [20] 从用户背景信息、微博文本、社交信息对关系强度进行研究 (TS-BPI). 其次, 为了分别说明本文引入共同节点连接边和社交历史习惯对关系强度预测的影响, 本文设计两个对比实验 (DSTS-ATI-1 方法和 DSTS-ATI-2 方法). 其中, DSTS-ATI-1 方法在计算用户关系强度时, 直接以共同邻居节点来衡量网络结构连接强度, 未考虑共同节点连接边数; DSTS-ATI-2 方法并未赋予社交行为不同的权重, 不考虑社交历史习惯对关系强度的影响.

4.4 预测结果评价与分析

4.4.1 参数设置

为了求得公式 (4) 中的系数取值范围, 本文提出了本文的实验算法, 如图 5 所示. 在满足 $\alpha+\beta+\gamma=1$, $0 \le \alpha \le 1$, $0 \le \beta \le 1$, $0 \le \gamma \le 1$ 的基本条件, 本文以 Top-N 用户关系强度列表评价标准, 在 $NDCG$ 取值最大的情况下, 得到了以下结果, 如表 1 所示.

观察图 6 可知, α 的取值主要集中在 $[0.20, 0.28]$ 之间, 上四分位数是 0.28, 中位数是 0.25, 下四分位数是 0.22; β 的取值主要在集中在 $[0.28, 0.32]$ 之间, 上四分位数是 0.32, 中位数是 0.29, 下四分位数是 0.28; γ 的取值主要集中在 $[0.42, 0.5]$ 之间, 上四分位数是 0.50, 中位数 0.47, 下四分位数是 0.43. 通过比较这三个系数平均值的取值范围可发现, 用户的社交行为是对用户关系强度最重要的影响因素, 特征属性和网络结构对用户关系强度影响相对较弱.

4.4.2 融合特征属性、网络结构、社交互动的微博用户关系强度计算模型预测说明

用户 U_1 的 Top-10 用户关系强度计算结果如表 2 所示. 观察表 2 可知, 交互双方对一段关系的评价是不对称的, 例如, 在用户 U_1 的标准 Top-10 用户关系强度列表中, 从用户 U_1 的角度感知, 用户 U_1 与用户 2 的关系强度较强; 相反, 从用户 2 的角度感知, 用户 2 与用户 U_1 的关系强度较弱, 本文的计算结果也验证了这一点. 这说明用户之间的关系强度是存在方向性差异, 从种子用户的角度感知到的关系强度和从交互对象角度感知关系强度存在一定差异.

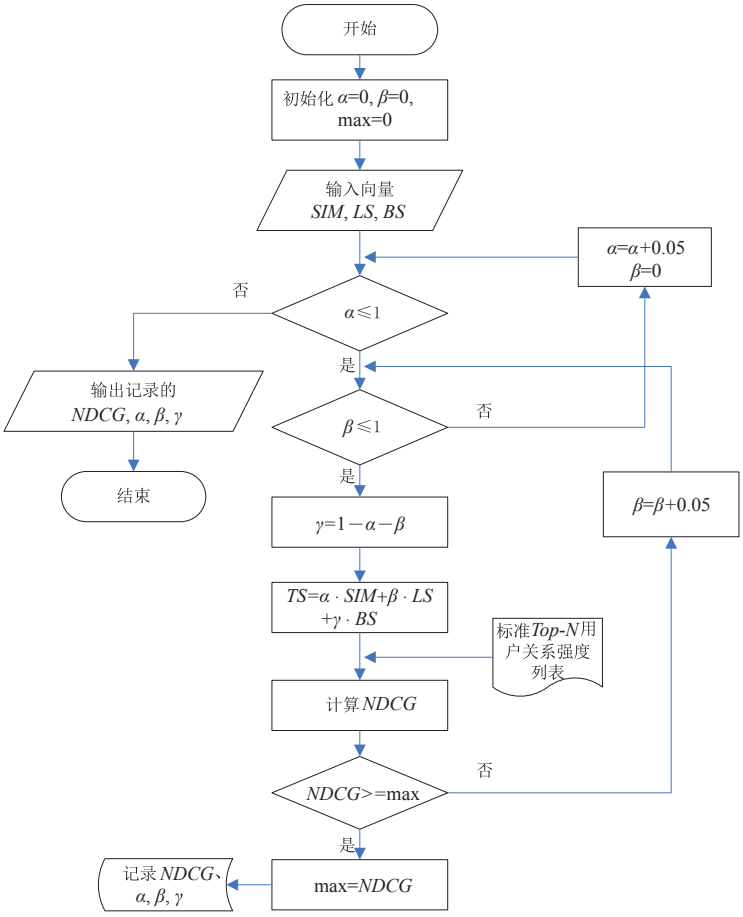

图 5 实验算法流程图

表 1 α、β、γ 参数取值箱图说明

变量名	最小值	下四分位数	中位数	上四分位数	最大值
α	0.10	0.22	0.25	0.28	0.30
β	0.16	0.28	0.29	0.32	0.35
γ	0.38	0.43	0.47	0.50	0.74

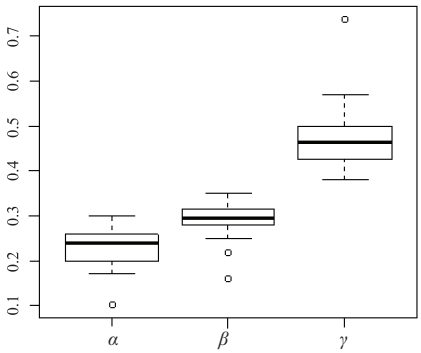

图 6 20 个种子用户 α、β、γ 参数取值箱图

表 2 用户 U_1 的 Top-10 用户关系强度计算结果说明

序号	单向 $U_1 \leftarrow$		单向 $U_1 \rightarrow$		标准排序 $U_1 \leftarrow$		标准排序 $U_1 \rightarrow$	
	交互对象	排名	交互对象	排名	交互对象	排名	交互对象	排名
1	用户 3	1	用户 3	2	用户 1	1	用户 1	2
2	用户 1	2	用户 1	1	用户 2	2	用户 2	8
3	用户 2	3	用户 2	10	用户 3	3	用户 3	1
4	用户 4	4	用户 4	3	用户 4	4	用户 4	3
5	用户 5	5	用户 5	6	用户 5	5	用户 5	4
6	用户 7	6	用户 7	NULL	用户 6	6	用户 6	5
7	用户 8	7	用户 8	7	用户 7	7	用户 7	NULL
8	用户 6	8	用户 6	4	用户 8	8	用户 8	7
9	用户 9	9	用户 9	NULL	用户 9	9	用户 9	NULL
10	用户 11	10	用户 11	9	用户 10	10	用户 10	9

由图 7 可知, 在研究方法上, 本文提出的 DSTS-ATI 方法在每个领域中 $NDCG$ 取值最高, TS-BP 方法 N-DCG 值最低, TS-I 方法和 TS-BPI 方法次之. 说明单一维度方法较多维度混合的方法而言, 对强关系用户的发现能力较弱. 影响社交网络用户关系强度的因素很多, 随着维度的增加, 综合考虑各维度上的因素能够增加预测准确性; 在不同领域上, 体育、电商领域的 $NDCG$ 值较高. 这主要是因为这两个领域种子用户社交圈相对较小, 对专业要求更高, 干扰因素较少, 区分更明显; 观察图 8 可知, 随着 N 的增加, $NDCG$ 取值增大, 发现强关系用户的能力增强.

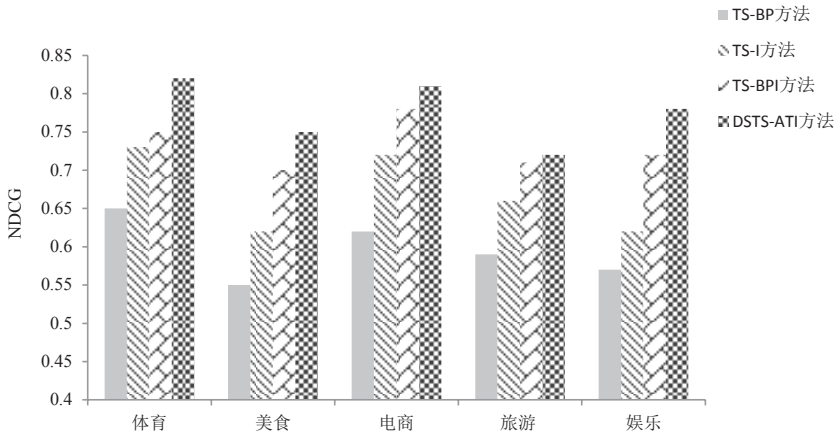

图 7　各方法计算得到的平均 NDCG

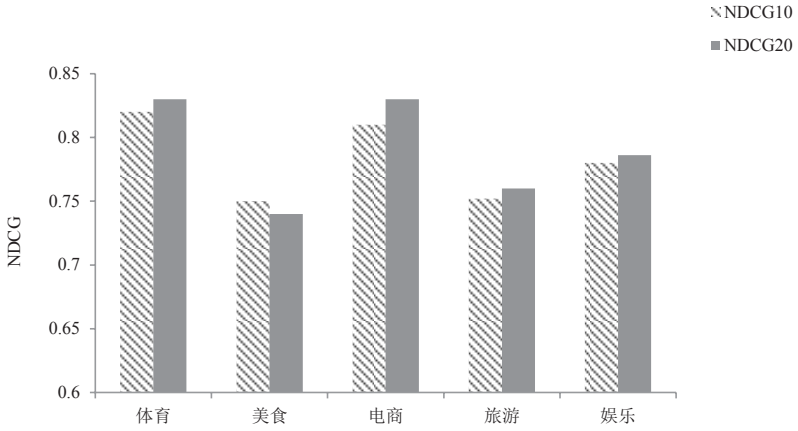

图 8　本文 DSTS-ATI 方法计算得到的平均 NDCG10、NDCG20

4.4.3　用户关系强度不对称性说明

体育、美食、电商、旅游、娱乐五个领域的用户关系网络局部示意图如图 9 所示. 其中, 黑点为领域内的种子用户, 白点为交互对象, 箭头表示用户之间存在关注行为, 图中数字为本文模型计算所得的用户关系强度值 (保留一位小数), 种子用户和交互对象之间的关注关系存在单向关注和双向关注, 且同一领域内的社交用户连接紧密, 不同领域内的社交用户连接稀疏. 通过观察图 9 数值和表 2 用户排名可发现, 交互双方对同一段关系的评价是不一样, 由于交互双方在同一关系中所处的地位不同, 因此他们所感知的关系强度是存在方向性差异的. 这主要是因为微博用户关系的建立和社交行为存在方向性, 微博用户关系强度的感知是不对称的. 此外, 图中 A、B、C 三点处入度高, 且感知的用户关系强度大, 因此, 这三点为改关系网络中影响力大的节点. 故此, 本文所提的融入网络结构与社交习惯的不对称用户关系强度计算方法对社交网络中的意见领袖的发现有重要意义.

4.4.4　λ_1、λ_2 参数对关系强度预测精度的影响

在计算用户关系强度的实验中, 公式 (4) 中的 λ_1、λ_2 系数取值会影响到预测结果的精确性. λ_1、λ_2 系数分别有 11 种取值情况, 具体为 $[0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0]$. 在满足 $\lambda_1 + \lambda_2 = 1$, $0 \le \lambda_1 \le 1$, $0 \le \lambda_2 \le 1$ 的基本条件下, 本文对取值范围进行了实验. 实验发现在 λ_1 取 0.4 时, λ_2 取 0.6 时, $NDCG$ 取值最大为 0.82, 如图 10 所示.

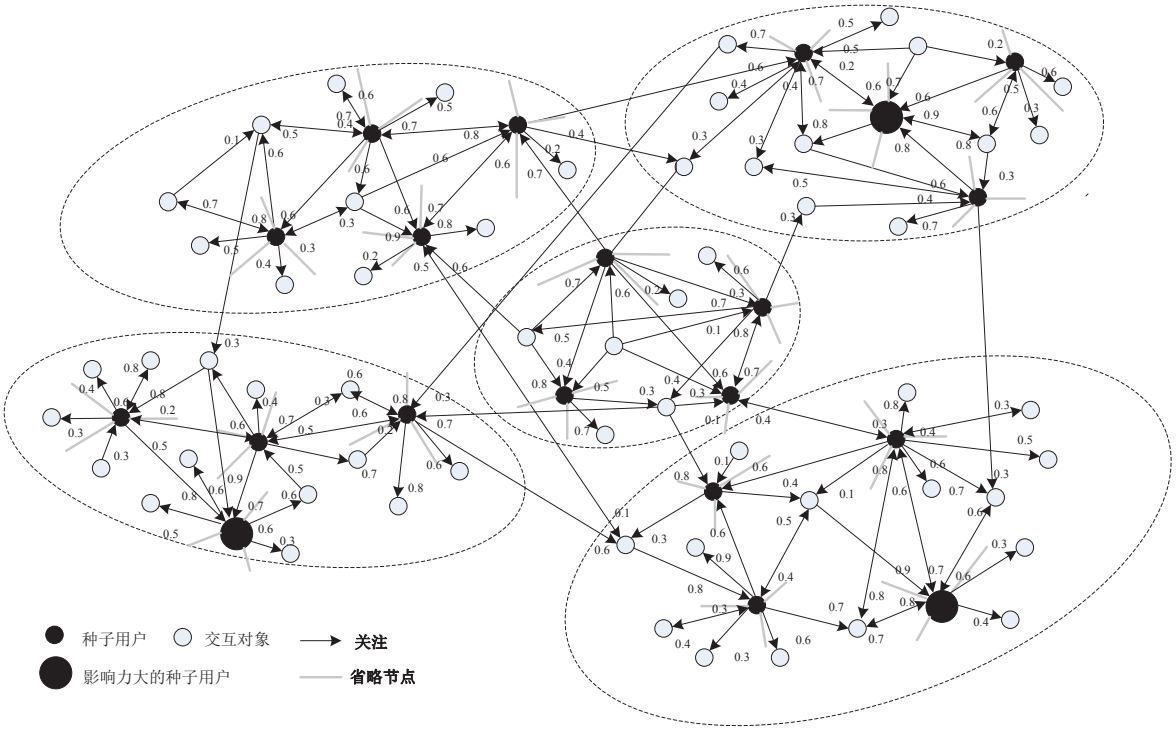

图 9 用户关系网络局部示意图

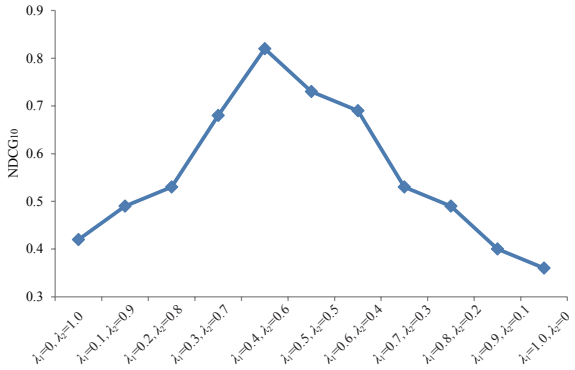

图 10 λ_1、λ_2 系数取值对关系强度预测精度的影响说明

4.4.5 共同邻居节点连接边对关系强度预测精度的影响

观察图 11 可知, 本文考虑共同节点连接边能够提高 NDCG 值. 进一步观察不同领域的增长幅度可发现, 在体育领域和电商领域增加幅度小, 在美食、旅游、娱乐领域增长的幅度大. 这主要是因为体育领域和电

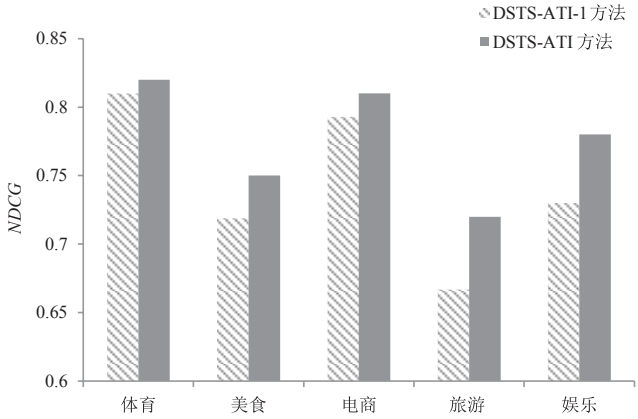

图 11 共同邻居节点连接边对关系强度预测精度的影响

商领域专业性强, 微博交互对象多为线下好友转化而来, 关系比较稳定, 因此, 共同好友连接边引入对用户关系强度影响小. 而美食、旅游、娱乐领域较其他两个领域而言, 专业性低, 微博交互对象多为有共同兴趣的陌生人, 关系不稳定, 因此, 引入共同邻居节点边数对强关系用户的发现能力影响大.

4.4.6 社交习惯对关系强度预测的影响

由图 12 可知, 引入用户社交历史习惯可以增强发现强关系用户的能力. 这主要是因为用户在社交活动中会表现出对某一社交行为或某些社交对象产生明显的依赖. 这种社交习惯使得用户社交活动的存在疏密, 会影响用户关系强度的感知, 因此在计算关系强度时不能把所有社交行为一概而论, 每种社交行为都应该赋予不同的权重.

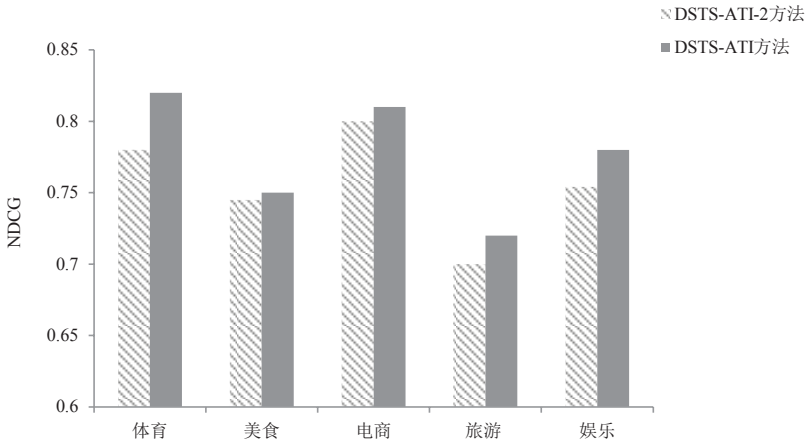

图 12　社交习惯对关系强度预测的影响

5 结束语

本文融合特征属性、网络结构、社交互动三个维度信息, 提出了不对称微博用户关系强度预测模型. 为了解决计算用户相似性时数据稀疏性问题, 综合考虑了用户的个人背景信息和微博短文本信息, 来计算用户属性特征相似度; 在计算用户网络结构连接强度时, 研究了共同邻居节点和共同邻居节点连结边数对关系强度的影响; 区别于无向关系网络, 微博用户能够单方向订阅感兴趣的用户, 与之建立关注和粉丝关系或者相互关注关系, 且在社交网络中用户的交互活动有一定的习惯性. 因此本文将用户社交习惯作为惩罚因子, 来计算该社交习惯对用户关系形成的贡献权重, 并且从交互双方分别衡量用户关系强度. 融入网络结构与社交习惯的不对称用户关系强度计算方法的提出为微博意见领袖的发现和信息传播机制的研究奠定了基础.

参考文献

[1] 李立耀, 孙鲁敬, 杨家海. 社交网络研究综述 [J]. 计算机科学, 2015, 42(11): 8–21.
Li L Y, Sun L J, Yang J H. Research on online social network[J]. Computer Science, 2015, 42(11): 8–21.

[2] Huberman B A, Romero D M, Wu F. Social networks that matter: Twitter under the microscope[J]. Social Science Electronic Publishing, 2009, 14(1): 115–121.

[3] 朱涵钰, 吴联仁, 吕廷杰. 社交网络用户隐私量化研究: 建模与实证分析 [J]. 清华大学学报 (自然科学版), 2014(3): 402–406.
Zhu H Y, Wu L R, Lü T J. Research on quantifying user privacy on social networking sites[J]. Journal of Tsinghua University (Science & Technology), 2014(3): 402–406.

[4] Granovetter M S. The strength of weak ties[J]. American Journal of Sociology, 1973, 78(6): 347–367.

[5] Granovetter M S. Getting a job: A study in contacts and careers[M]. 2nd ed. Chicago: University of Chicago Press, 1994.

[6] Koo D M. Impact of tie strength and experience on the effectiveness of online service recommendations[J]. Electronic Commerce Research & Applications, 2016, 15: 38–51.

[7] Frenzen J, Nakamoto K. Structure, cooperation, and the flow of market information[J]. Journal of Consumer Research, 1993, 20(3): 360–75.

[8] Wellman B, Wortley S. Different strokes for different folks: Community ties and social support[J]. American Journal of Sociology, 1990, 96(3): 558–588.

[9] Krackhardt D. The strength of strong ties: The importance of philos in organizations[J]. Networks & Organizations, 1992: 216–239.
[10] Sett N, Ranbir Singh S, Nandi S. Influence of edge weight on node proximity based link prediction methods[J]. Neurocomputing, 2015, 172(C): 71–83.
[11] Zuo X, Blackburn J, Kourtellis N, et al. The power of indirect ties[J]. Computer Communications, 2015, 73(PB): 188–199.
[12] Shen C C, Chiou J S, Hsiao C H, et al. Effective marketing communication via social networking site: The moderating role of the social tie [J]. Journal of Business Research, 2016, 69(6): 2265–2270.
[13] Wang J C, Chang C H. How online social ties and product-related risks influence purchase intentions: A Facebook experiment[J]. Electronic Commerce Research & Applications, 2013, 12(5): 337–346.
[14] Wei J, Bing B, Guo X, et al. The process of crisis information dissemination: Impacts of the strength of ties in social networks[J]. Kybernetes, 2014, 43(2): 178–191.
[15] Kahanda I, Neville J. Using transactional information to predict link strength in online social networks[C]// International Conference on Weblogs and Social Media, California, 2009: 235–243
[16] Luarn P, Chiu Y P. Key variables to predict tie strength on social network sites[J]. Internet Research, 2015, 25(2): 218–238.
[17] 邓晓懿, 金淳, 韩庆平, 等. 基于情境聚类和用户评级的协同过滤推荐模型 [J]. 系统工程理论与实践, 2013, 33(11): 2945–2953.
Deng X Y, Jin C, Han Q P, et al. Improved collaborative filtering model based on context clustering and user ranking[J]. Systems Engineering — Theory & Practice, 2013, 33(11): 2945–2953.
[18] 仲兆满, 胡云, 李存华, 等. 微博中特定用户的相似用户发现方法 [J]. 计算机学报, 2016, 39(4): 765–779.
Zhong Z M, Hu Y, Li C H, et al. Discovering similar users for specific on microblog[J]. Chinese Journal of Computers, 2016, 39(4): 765–779.
[19] Xiong L, Lei Y, Huang W, et al. An estimation model for social relationship strength based on users' profiles, co-occurrence and interaction activities[J]. Neurocomputing, 2016, 214: 927–934.
[20] 徐志明, 李栋, 刘挺, 等. 微博用户的相似性度量及其应用 [J]. 计算机学报, 2014, 37(1): 207–218.
Xu Z M, Ll D, Liu T, et al. Measuring similarity between microblog users and its application[J]. Chinese Journal of Computers, 2014, 37(1): 207–218.
[21] Tang X, Miao Q, Quan Y, et al. Predicting individual retweet behavior by user similarity[J]. Knowledge-Based Systems, 2015, 89(C): 681–688.
[22] Liu Z Y, Chen X X, Sun M S. A simple word trigger method for social tag suggestion[C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2011: 1577–1588.
[23] Chen J, Liu Y, Zou M. Home location profiling for users in social media[J]. Information & Management, 2016, 53(1): 135–143.
[24] Burt R S. Structural holes: The social structure of competition[M]. Harvard University Press, 2010.
[25] Chulyadyo R, Leray P. A personalized recommender system from probabilistic relational model and users' preferences[J]. Procedia Computer Science, 2014, 35: 1063–1072.
[26] Cannistraci C V, Alanis-Lobato G, Ravasi T. From link-prediction in brain connectomes and protein interactomes to the local-community-paradigm in complex networks[J]. Scientific Reports, 2015, 3(4): 1613.
[27] Gilbert E, Karahalios K. Predicting tie strength with social media[C]// Sigchi Conference on Human Factors in Computing Systems. ACM, 2009: 211–220.
[28] Arnaboldi V, Guazzini A, Passarella A. Egocentric online social networks: Analysis of key features and prediction of tie strength in Facebook[J]. Computer Communications, 2013, 36(10–11): 1130–1144.
[29] Qi G J, Aggarwal C C, Huang T. Community detection with edge content in social media networks[C]// International Conference on Data Engineering. IEEE, 2012: 534–545.
[30] Ma H, Zhou D, Liu C, et al. Recommender systems with social regularization[C]// Forth International Conference on Web Search and Web Data Mining, WSDM 2011, Hong Kong, China, 2011: 287–296.
[31] Watts D J, Strogatz S H. Collective dynamics of 'small-world' networks[J]. Nature, 1998, 393(6684): 440–442.
[32] Milgram S. The small world problem[J]. Psychology Today, 1967, 2(1): 185–195.
[33] 白林根, 谌志群, 王荣波, 等. 微博关注关系网络 K- 核结构实证分析 [J]. 现代图书情报技术, 2013(11): 68–74.
Bai L G, Chen Z Q, Wang R B, et al. Empirical analysis on K-core microblog follwering relationship network[J]. New Technology of Library and Information Service, 2013(11): 68–74.
[34] 于岩, 陈鸿昶, 于洪涛. 基于霍克斯过程的社交网络用户关系强度模型 [J]. 电子学报, 2016, 44(6): 1362–1368.
Yu Y, Chen H C, Yu H T. A social network user relationship strength model based on Hawkes process[J]. Acta Electroncia Sinica, 2016, 44(6): 1362–1368.