

社会化问答网站知识传播网络推断方法

左 遥^{1),2)} 梁 英¹⁾ 毕晓迪^{1),2)} 史红周¹⁾ 董祥祥^{1),2)} 李锦涛¹⁾

¹⁾(中国科学院计算技术研究所 北京 100190)

²⁾(中国科学院大学 北京 100190)

摘 要 社会化问答网站通过用户间提问与回答的方式帮助用户获得全面的知识. 知识传播过程是进一步研究信息传播规律和社交行为的基础, 由于无从得知传播者影响了哪些用户, 因而无法直接获得答案的完整传播过程. 该文提出了一种描述社会化问答网站知识传播过程的传播网络模型, 定义了社会化问答网站知识从传播者到接收者的传播过程, 归纳出 follow, notification, qlink, recommendation 这 4 种可能的知识传播方式, 给出了知识传播网络定义, 为传播网络推断奠定了理论基础. 提出了一种社会化问答网站知识传播网络推断方法, 分别提取节点类型、节点时间关系及候选传播者排序特征推断 follow 关系, 采用启发式方法推断另外 3 种关系, 生成知识传播网络. 结合知乎问答数据对所提方法进行了实验验证及实例分析, 结果表明推断方法有效并具有较高的准确率.

关键词 社会化问答网站; 知识传播方式; 信息传播模型; 传播网络推断; 知识传播过程

中图法分类号 TP311

DOI 号 10.11897/SP.J.1016.2018.00082

An Inference Method of Knowledge Diffusion Network in Community Question Answering Sites

ZUO Yao^{1),2)} LIANG Ying¹⁾ BI Xiao-Di^{1),2)} SHI Hong-Zhou¹⁾
DONG Xiang-Xiang^{1),2)} LI Jin-Tao¹⁾

¹⁾(Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190)

²⁾(University of Chinese Academy of Sciences, Beijing 100190)

Abstract With the rise of Web2.0, many community question answering (CQA) websites have emerged. Community question answering combines the advantages of social networking websites, Wikipedia and traditional Q&A websites, it has three main functions (follow topic, follow question, follow user) to present the best answer and comprehensive knowledge, as well as build social relationships among users. With the opening of registration, the original elite community becomes much more popular. In order to guide the behavior in the process of information diffusion, maintain high content quality when users have exploded and achieve better development, we need to analyze the process of information diffusion on CQA websites, explore diffusion peculiarities and internal rules. Current research of CQA websites is limited to analyzing raising and answering questions, such as finding experts for certain questions, predicting questioner satisfaction and evaluating answer quality, etc. At the same time, current research for CQA websites lacks attention for some hot research topics in social networking sites, such as how friend relationship influence information diffusion, how new messages spread among the network, and predicting hotness of new messages. The main reason is that social networking site records every sender of information propagation, while CQA website only records the overall situation, making it hard to measure each user's influence. As a result, the diffusion process of each answer cannot be

收稿日期:2016-09-01;在线出版日期:2017-05-26. 本课题得到国家重点研发计划(2016YFB0800403)、国家“八六三”高技术研究发展计划项目基金(2015AA015803)、北京市科技计划课题(Z161100001616009)资助. 左 遥,男,1991年生,硕士,主要研究方向为大数据分析 & 挖掘. E-mail: laike9m@gmail.com. 梁 英(通信作者),女,1962年生,硕士,高级工程师,中国计算机学会(CCF)高级会员,主要研究方向为大规模数据挖掘与存储处理、泛在计算、信息安全. E-mail: liangy@ict.ac.cn. 毕晓迪,女,1992年生,硕士研究生,中国计算机学会(CCF)学生会员,主要研究方向为大数据分析 & 挖掘. 史红周,男,1971年生,博士,高级工程师,中国计算机学会(CCF)会员,主要研究方向为嵌入式系统、泛在计算、物联网安全. 董祥祥,男,1996年生,硕士研究生,主要研究方向为大数据分析 & 挖掘. 李锦涛,男,1962年生,博士,研究员,中国计算机学会(CCF)高级会员,主要研究领域为社会化媒体计算、多媒体内容分析和网络计算技术等.

obtained directly from CQA, which is the theoretical basis for further exploration of information diffusion law and social behaviors in CQA. Facing these challenges, we study the key technologies relating to CQA websites knowledge diffusion process, including knowledge diffusion modeling in CQA website and knowledge diffusion process inference. We use Zhihu as a data source to verify the effectiveness of our methods. Our main contributions are as follows: (1) We define the knowledge diffusion process from sender to receiver in CQA and propose a diffusion network model to describe it, landing theoretical foundation for diffusion network inference. Starting from the concepts in CQA website, we analyze their relations, define knowledge diffusion process in CQA website and sum up four knowledge transmission modes, including follow, notification, qlink and recommendation. We then introduce priority assumption and first activated independent cascade model assumption. On this basis, the knowledge diffusion network definition is given. (2) We propose a method to infer knowledge diffusion network. Experiment result proves the effectiveness of our method. Our knowledge inference method uses data from stable answers as input, it does not depend on time-slicing data, thus reduces the time of data acquisition. Given that input data lacks accurate time of upvoters and question followers, we treat the inference of follow relation as a classification problem, extracting node type, node time relation and candidate sender relative order as features, and use machine learning techniques to do classification. Then we use a heuristic method to infer other three kinds of relations and generate the whole knowledge diffusion network. Finally, we conduct experiments on question & answer data collected from Zhihu and analyze the results. The experiment results show that our inference method is effective and has high accuracy.

Keywords community question answering; knowledge transmission mode; information diffusion model; diffusion network inference; knowledge diffusion process

1 引言

社会化问答网站随着 Web2.0 时代的兴起不断涌现,人们对社会化问答网站的关注也逐步提升^[1]. 社会化问答网站融合了社交网站、百科类网站和传统问答网站的优点,它通过三大功能(关注话题、关注问题、关注用户)来呈现问题回答的最佳答案和全面的相关知识.随着注册的开放,作为精英社区的社会化问答网站逐渐大众化.为引导网络信息传播过程中的传播行为,帮助社会化问答网站在用户数暴增时保持内容优质,实现更好地发展,需针对社会化问答网站的知识传播过程进行分析,发现传播过程的特点及内部规律,为进一步深入探索社会化问答网站提供一定的理论研究基础.开展社会化问答网站知识传播网络推断方法研究,对社会化问答网站的拓展应用及推动社会化问答网站的发展具有重大意义.

完整的信息传播过程之所以重要,是因为它是进一步研究信息传播规律和社交行为的基础,比如好友关系网络结构对于信息传播的影响、消息的扩散规律以及转发数预测等,都依赖于传播网络.目前这些研究一般以社交网络为对象,极少涉及社会化问答网站.下面以社交网络的微博和社会化问答网站的知乎为例,对比社交网络和社会化问答网站的

信息传播过程.微博通过“转发数”记录了直接转发该博文的每个用户,因此不论博文被转发多少次,总能根据每一次的转发记录还原一条微博从发博文的用户到达转发了这条博文的所有用户的完整传播过程,因此传播过程数据能从社交网站直接获得.而知乎通过“点赞数”记录了所有直接或间接看到某问题一个答案的所有用户,因此无法从社会化问答网站上得知一个传播者直接影响了哪些用户,也无法得知一个用户是从哪个传播者接收到了信息,因而无法直接获得知识(答案)的完整传播过程.

当前针对社会化问答网站的研究局限于问答行为,比如分析问题、答案的质量,预测提问者满意度,这些研究往往以问答文本和用户历史行为作为依据,缺乏对于传播过程信息的利用,导致研究无法深入.研究面临的困境迫切地需要我们找到一种方法来获取社会化问答网站中的知识传播过程,这为我们提出了 2 项挑战:(1)如何描述社会化问答网站中的知识传播过程;(2)如何推断社会化问答网站中的知识传播过程.

本文针对社会化问答网站知识传播过程的推断方法进行了探索,总结了 4 种可能的知识传播方式,定义了社会化问答网站知识传播网络及其相关操作,使用 SVM 和启发式方法分别进行不同类型的传播关系推断,生成知识传播网络.

2 相关工作

当前对社会化问答网站的研究主要集中于问答行为、专家发现和社交关系. 问答行为包括问答质量分析和提问者满意度预测. Li 等人^[2]、Zhou 等人^[3]分别使用 MRLP 算法和 LSTM 算法预测问题质量. Liu 等人^[4]基于 Yahoo!Answers 定义提问者满意度, 使用多种分类算法进行满意度预测. Shah^[5]发现 Yahoo!Answers 中较晚回答问题的回答者给出的答案质量更高. 专家发现是指为问题推荐最适合的回答者, 或发现在某一方面具有权威性的用户. 传统的专家发现方法可分为用户描述方法和问题相似性方法^[6]. 用户描述方法根据用户历史回答推测用户擅长的主题和兴趣; 问题相似性方法寻找相似问题中提供了排名靠前答案的用户. Liu 等人^[7]提出了一种对话题敏感的专家推荐算法 ZhihuRank, 试图尽量减少用户人气对于用户在某问题下权威程度的干扰. Zhou 等人^[8]将专家发现看作二分类问题, 使用 SVM 模型进行分类. 社会化问答网站在传统问答网站的基础上加入了社交功能, 社交关系对于用户的问答行为带来了诸多影响. Wang 等人^[9]研究了社交关系对 Quora 中的问题、回答、赞同的影响. Rodrigues 等人^[10]把用户使用社会化问答网站的目的分为社交和分享知识 2 类, 并使用分类算法揭示用户参与社会化问答网站的价值取向. 目前关于社会化问答网站的研究更多集中于研究问答行为, 即使涉及社交行为, 也更多地是将用户关注关系作为一种特征, 而缺乏对社会化问答网站知识传播的深入研究, 其主要原因是社会化问答网站没有直接可以利用的信息传播过程数据.

社会化问答网站虽然在内容方面和传统问答网站差异不大, 但在形式上更接近社交网站, 社会化问答网站和社交网站组织结构的相似性直接导致了信息在两类网站上的传播过程相似性. 链路预测是社交网站信息传播研究中的一项重要内容, 它通过已知的网络节点以及网络结构等信息预测网络中尚未产生连边的 2 个节点之间产生链接的可能性. 这种预测方法既包含了对未知链接的预测, 也包含了对未来链接的预测^[11], 其本质是探究网络演化规律和网络内部机制^[12], 包括在已知部分节点类型的网络中预测未标注节点的类型.

链路预测方法主要有 2 类, 分别是基于相似性的链路预测和基于最大似然的链路预测. 最简单的方式是衡量节点属性与网络结构, 判断 2 个节点的共同邻居, 以多种方式计算相似度. 在此基础上, Ou 等人^[13]提出了一种基于网络资源分配的相似度指标, 假设 2 个节点的共同邻居可以用来传递资源, 则

从节点 A 传递到节点 B 的资源量就代表了 A 与 B 的相似度. 但这种衡量方式对于网络聚集度有较大的依赖, 不具有普遍性. 若网络具有明显的层次组织, 则基于最大似然的链路预测会有更高的精确度. Clauset 等人^[14]用树形图表示网络的层次结构, 给每个内部节点一个概率值, 2 个节点连接的概率等于他们的最近公共祖先节点的概率, 并根据树形图定义网络的似然估计值. 但此类方法复杂度较高, 不适合较大规模网络.

虽然社会化问答网站与社交网站组织结构相似, 但节点属性存在差异, 故仅依靠社会化问答网站节点间相似性不能准确推断知识传播过程. 本文对社会化问答网站可能的知识传播方式进行归纳总结, 根据节点行为特征, 判断节点之间时间先后关系, 生成候选边, 分析传播方式求解传播路径, 提出了适用于不同传播方式的知识传播过程推断方法, 并以知乎为例进行了实例分析和验证. 本文的主要贡献为: (1) 归纳了 4 种社会化问答网站知识传播方式 follow, notification, qlink, recommendation; (2) 提出了一种社会化问答网站知识传播过程的传播网络模型, 为传播网络推断奠定了理论基础; (3) 提出了一种社会化问答网站知识传播网络推断方法, 生成知识传播网络.

3 社会化问答网站知识传播模型

3.1 社会化问答网站相关概念

3.1.1 基本元素

问题. 社会化问答网站的基本单元, 记作 Q . 每个问题有唯一 ID, 记作 qid .

答案. 用户对于问题给出的解答, 记作 A . 每个答案有唯一 ID, 记作 $aid(A)$, 在不引起歧义的情况下也可直接写作 aid . 记问题 Q 的对应的答案的集合为 $Ans(Q)$.

本文中的知识传播特指问题的答案传播.

图 1 展示了本文涉及的以问题和答案为中心的

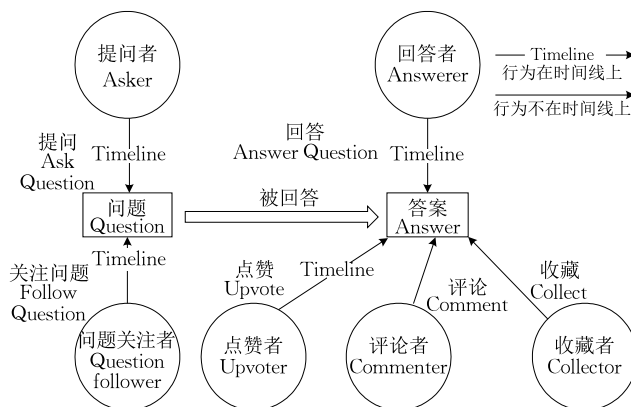


图 1 社会化问答网站相关概念及关系

社会化问答网站相关概念及其之间的关系,下文将逐一进行介绍。

3.1.2 用户类型

记社会化问答网站用户的集合为 $U = \{u\}$, 每个用户 u 有唯一 ID, 记作 uid . 本文中, u 和 uid 一一对应, 对 u 和 uid 不作区分, 根据上下文不同, 可以混用. 本文根据问题、答案、用户关系和用户行为定义了不同场景下的用户类型。

(1) 问题场景

提问者. 提出某个问题的用户, 问题 Q 的提问者记作 $asker(Q)$, $asker(Q) \in U$.

问题关注者. 关注某个问题的用户, 问题 Q 的关注者集合记作 $Follower(Q)$, $Follower(Q) \subset U$.

(2) 答案场景

回答者. 提交问题 Q 的答案 A 的用户, 记作 $answerer(A)$, $answerer(A) \in U$. 一个问题 Q 的回答者集合记作 $Answerer(Q)$, $Answerer(Q) \subset U$.

点赞者. 给答案点赞的用户, 答案 A 的点赞者集合记作 $Upvoter(A)$, $Upvoter(A) \subset U$.

评论者. 在答案下留下评论的用户, 答案 A 的评论者集合记作 $Commenter(A)$, $Commenter(A) \subset U$.

收藏者. 收藏答案的用户, 答案 A 的收藏者集合记作 $Collector(A)$, $Collector(A) \subset U$.

(3) 用户关系场景

关注者. 关注了某用户 u 的用户, 记作 $follower$, 关注者集合记作 $Follower(u)$, $Follower(u) \subset U$.

被关注者. 被用户 u 关注的用户, 记作 $followee$, 被关注者集合记作 $Followee(u)$, $Followee(u) \subset U$.

3.1.3 用户行为

提问. 用户提出问题 Q 的行为, 记作 Act_{ask} , 该用户同时成为提问者 $asker(Q)$.

回答. 用户撰写并提交答案 A 的行为, 记作 Act_{answer} , 该用户同时成为回答者 $answerer(A)$.

关注问题. 用户通过点击“关注问题”按钮关注问题 Q 的行为, 简记作 $Act_{followQ}$. 该用户同时成为问题关注者 $Follower(Q)$ 之一。

点赞. 用户通过社交网络的“赞”功能对答案 A 表示“赞同”、“喜爱”、“点赞”的行为, 记作 Act_{upvote} , 该用户同时成为点赞者 $Upvoter(A)$ 之一。

评论. 用户评论答案 A 的行为, 记作 $Act_{comment}$, 该用户同时成为评论者 $Commenter(A)$ 之一。

收藏. 用户收藏答案 A 的行为, 记作 $Act_{collect}$, 该用户同时成为收藏者 $Collector(A)$ 之一。

关注用户. 用户关注另一个用户的行为, 记作 $Act_{followU}$.

在实际的应用中, 还存在“感谢”、“反对”等用户行为, 由于这些用户行为无法统计, 并且不会直接影响传播网络推断, 因此不在本文研究范围之内。

下面给出用户行为的形式化定义。

定义 1. 用户行为($Action$).

$$Action = \langle uid, acttype, time, item_id \rangle \quad (1)$$

式(1)中:

uid 是触发行为的用户的 ID;

$acttype$ 是用户行为类型,

$$acttype \in \{Act_{ask}, Act_{answer}, Act_{followQ}, Act_{upvote}, Act_{comment}, Act_{collect}, Act_{followU}\};$$

$time$ 是用户行为发生的时间;

$item_id$ 是可选属性, 表示与该用户行为对应的社会化问答网站的基本元素, 即答案 aid 或问题 qid ,

$$item_id = \begin{cases} aid, & acttype \in \{Act_{answer}, Act_{upvote}, \\ & Act_{comment}, Act_{collect}\} \\ qid, & acttype \in \{Act_{ask}, Act_{followQ}\} \\ \emptyset, & acttype \in \{Act_{followU}\} \end{cases}.$$

设所有用户行为的集合为 $ACTION$, $ACTION = \{Action_i | i = 1, 2, \dots, n_{act}\}$, 其中: $Action_i$ 表示社会化问答网站用户的行为, n_{act} 为行为数。

3.1.4 系统行为

系统推荐. 社会化问答网站给用户推荐的答案. 推荐一般基于网站内部数据, 因此无法预测哪些答案会被推荐。

新答案提醒. 用户关注问题后, 每当新答案产生时收到的系统发出的通知, 通知中包含跳转到新答案的链接。

3.1.5 时间线

时间线. 出现在用户主页上的用户行为的列表. 用户 u 的时间线定义如下:

$$Timeline_u = [Action_1, Action_2, \dots, Action_i, \dots, Action_n] \quad (2)$$

式(2)中,

$$Action_i[acttype] \in \{Act_{ask}, Act_{answer}, Act_{followQ}, Act_{upvote}\},$$

$1 \leq i \leq n$, $n \in N^+$ 满足时间有序性,

$$Action_i[time] > Action_{i+1}[time], 1 \leq i \leq n-1.$$

只有提问、回答、关注问题和点赞行为会出现在时间线。

3.2 社会化问答网站知识传播方式

3.2.1 知识传播过程

下面针对社会化问答网站中答案传播过程中的关键概念给出定义。

定义 2. 传播者($Sender$): 指传播信息的用户. 对于社会化问答网站而言, 用户通过使自己的行为出现在其它用户的时间线上完成信息传播. 提问者、问题关注者、回答者、点赞者是潜在的信息传播者。

定义 3. 接收者($Receiver$): 指接收到信息的用户. 本文主要关注答案的传播, 对于答案 A , 其接收者 $Receiver$ 定义为

$$Receiver(A) = \{user | user \in Upvoter(A) \cup Commenter(A) \cup Collector(A)\} \quad (3)$$

准确判断一个用户是否接收(看)到了某个回答,除非获得网站内部数据,否则无法做到. 本文取真实答案接收者的一个子集,即所有和答案有交互行为的用户,作为答案接收者,包括点赞者 $Upvoter(A)$ 、评论者 $Commenter(A)$ 、收藏者 $Collector(A)$.

定义 4. 知识传播过程. 一个答案从回答者通过传播者传播并最终到达接收者的过程.

图 2 给出了知识传播过程的示意图. 在知识传播过程中,除回答者以外的传播者也是接收者,而接收者中的点赞者也可以是传播者.



图 2 知识传播过程

3.2.2 知识传播的 4 种方式

本文通过分析社会化问答网站的用户行为及答案传播特点,总结出社会化问答网站中答案由传播者传播到接收者的 4 种方式(如图 3 所示):follow 方式、notification 方式、qlink 方式和 recommendation 方式,为知识传播网络的关系构建提供依据.

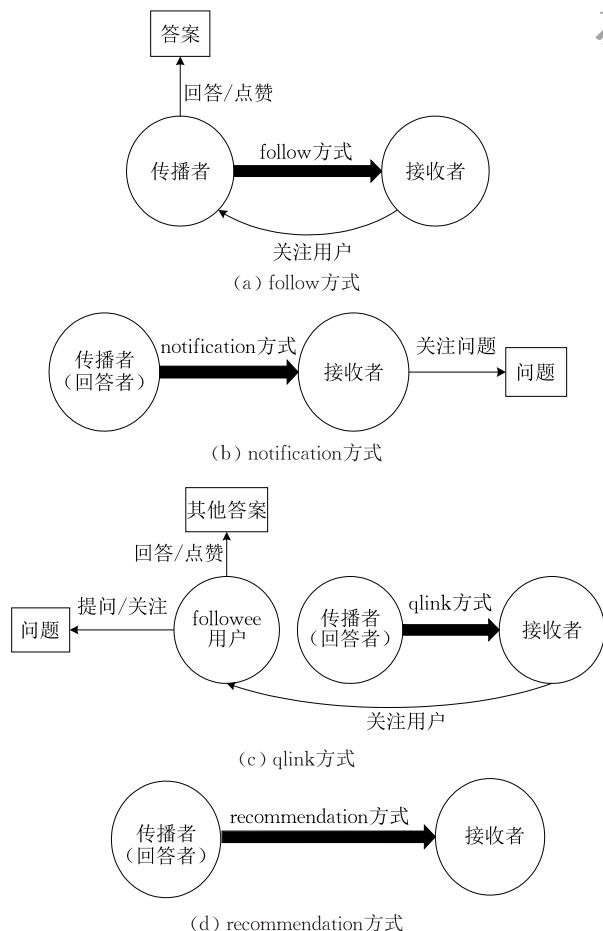


图 3 社会化问答网站的 4 种传播方式

(1) follow 方式. 接收者通过 followee 回答问题或对答案点赞直接看到答案.

(2) notification 方式. 问题关注者收到新答案提醒,通过点击新答案提醒中的链接跳转并看到答案,本文认为 notification 方式的传播者是回答者.

(3) qlink(question link)方式. 用户通过点进问题链接间接看到答案. 一般而言, qlink 包含 3 种典型方式:看到 followee 提问,看到 followee 关注问题,看到 followee 回答或点赞其它答案. 本文认为 qlink 方式的传播者是回答者.

(4) recommendation 方式. 除 follow, qlink, notification 这 3 种方式外其它看到答案的方式,如系统推荐. 同样地,认为 recommendation 方式的传播者也是回答者.

3.2.3 优先级假设

优先级假设对 4 种传播方式的可能性进行排序,当 2 种或多种传播方式并存时,假设用户通过高优先级的传播方式接收到答案. 优先级假设的依据是用户常规浏览习惯,即用户看到答案所需要耗费操作的多少影响用户看到答案的概率. 例如,若用户不需要任何操作就能看到答案,显然比用鼠标点击 2 次链接才能看到答案的可能性更大.

我们归纳总结了知乎网的特点以及用户的使用习惯,发现在时间线上能够直接看见由 follow 方式传播过来的答案,不需要任何操作,故其优先级最高. notification 方式需要用户点击新答案提醒中的链接,需要一次鼠标点击操作,优先级低于 follow 方式. qlink 方式需要用户首先点击问题链接进入问题,然后浏览其中的答案,花费 1 次点击操作以及若干次页面滚动操作,优先级低于 notification. recommendation 方式由定义可知其优先级是最低的. 因此有以下假设:

假设 1. 传播方式优先级假设. 4 种传播方式 follow, notification, qlink, recommendation 的优先级排序如下:

$$Priority_{follow} > Priority_{notification} > Priority_{qlink} > Priority_{recommendation} \quad (4)$$

第 4 节的传播网络推断算法遵循此优先级假设.

3.2.4 首次激活独立级联模型假设

信息传播模型描述了信息的传播方式. 经典的传播模型包括独立级联模型^[15]、创新扩散模型^[16]和传染病模型^[17]. 社会化问答网站中的知识传播过程中,每个用户的状态包含“接收”、“未接收”2 种,且接收到答案的用户才能传播给未接收的用户,更适合用独立级联模型进行建模.

信息级联是指信息或决策在一组个体中扩散

的过程. 满足 2 个条件: (1) 个体通过网络相连接; (2) 个体只能观察到其近邻的决策行为. 上文提出的 4 种传播方式, follow 和 qlink 方式中, 用户通过观察其 followee 的行为接收到答案, 用户通过好友关系网络相连. 我们把社会化问答网站的系统行为 (系统推荐、新答案提醒) 视为系统用户的行为, 并认为系统用户是所有其它用户的 followee, 则社会化问答网站中的知识传播过程符合独立级联模型假设, 即节点代表用户, 邻居节点指用户的 follower, 且用户处于活跃状态代表用户是答案的接收者, 激活操作指答案传播至当前不是接收者的用户并使该用户成为接收者. 本研究认为接收者一定是在答案首次传播过来时被激活的, 因此提出假设.

假设 2. 首次激活独立级联模型假设. 是一个扩展的独立级联模型, 满足以下 5 个条件:

- (1) 答案传播用有向图表述, 图中的节点表示社会化问答网站的用户; 一个节点只能影响其 follower.
- (2) 用户可以选择成为答案的接收者, 或者无视这个答案.
- (3) 一个用户一旦成为答案的接收者, 且身份是提问者、问题关注者、回答者、点赞者之一, 便有可能把答案传播给其 follower.
- (4) 用户只能从答案的非接收者变成答案的接收者.
- (5) 接收者是在答案首次传播到时成为接收者的.

3.3 社会化问答网站知识传播网络

3.3.1 知识传播网络定义

本节将从 1 个例子出发介绍知识传播模型, 图 4 显示了一个答案的传播过程, 实线表示知识传播方式, 虚线表示用户关注关系.

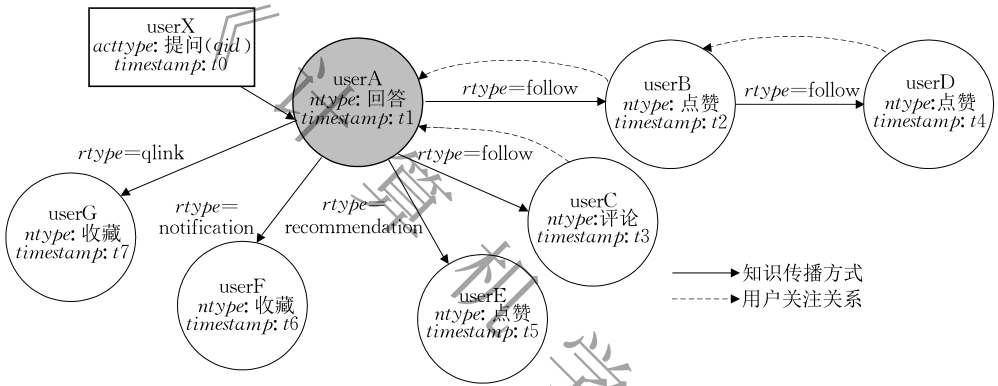


图 4 知识传播过程示例

userX 在 t_0 时刻提出了一个问题 (qid), userA 在 t_1 时刻回答了此问题 (aid). userB 关注了 userA, 因此看见 userA 提交了 1 个回答, userB 在 t_2 时刻给这个回答点赞. userD 关注了 userB, 看见 userB 对 userA 的答案点赞, 因此, userD 在 t_4 时刻也对 userA 的回答点赞. userC 亦关注了 userA, 在 t_3 时刻对回答进行了评论. userE 通过知乎的回答推荐看到了 userA 的回答, 在 t_5 时刻点赞. userF 在 userA 回答之前就关注了问题, 因此 userA 回答时, userF 收到了新答案提醒, 并在 t_6 时刻收藏. userG 在看到此问题后, 点问题链接看到了 userA 的回答, 在 t_7 时刻收藏.

由该传播实例可以看出, 知识传播是从传播者到接收者的一个过程, 传播过程形成了一个网络, 参与传播和接收信息的用户构成网络上的节点, 4 种传播方式描述节点之间的关系. 下面给出知识传播网络的定义.

定义 5. 知识传播网络 (G). 一个问题 Q 的一个答案 A 的知识传播网络是一个有向无环图 G_A :

$$G_A = (V, E | node \in V, relation \in E) \quad (5)$$

(1) V 是图中所有节点 ($node$) 的集合, 其中每个 $node$ 定义为

$$node = \langle uid, ntype, time \rangle \quad (6)$$

式 (6) 中 uid 表示用户 ID, 每个节点的 uid 唯一; $ntype$ 表示节点类型, 描述用户参与知识传播的行为, 取值集合为 $\{Act_{answer}, Act_{upvote}, Act_{comment}, Act_{collect}\}$ 的非空子集; $time$ 表示用户参与知识传播的时间. 每个 $node$ 在物理意义上均表示答案传播过程中经过的社会化问答网站用户, 由回答者和所有接收者组成.

(2) E 是图中所有关系 ($relation$) 的集合: 图 G_A 中从 $node_{src}$ 到 $node_{dest}$ 的边表示知识在 2 个用户之间传播的方式, 定义为

$$relation = \langle node_{src}, node_{dest}, rtype \rangle \quad (7)$$

式 (7) 中, $node_{src}$ 代表传播起始节点, 即传播者; $node_{dest}$ 代表传播终止节点, 即接收者; $rtype$ 为关系类型, 取值集合为 $\{follow, notification, qlink, recommendation\}$. 以下将 $relation$ 简写为 rel , 并使用 $rel[node][attr]$ 表示取 rel 中 $node$ 节点的 $attr$ 属性. 如 $rel[node_{dest}][uid]$ 表示提取 rel 关系中的传播终止节点 $node_{dest}$ 的用户 uid 属性.

follow 关系指答案通过 follow 方式由传播者 $node_{src}$ 传播到接收者 $node_{dest}$, 且满足式(8),

$$\begin{aligned} & \forall rel \in E, rel[rtype] = follow \rightarrow \\ & (\exists x \in node_{src}[ntype], x = Act_{answer} \vee x = Act_{upvote}) \wedge \\ & rel[node_{dest}][uid] \in Follower(rel[node_{src}][uid]) \wedge \\ & rel[node_{src}][time] < rel[node_{dest}][time] \end{aligned} \quad (8)$$

notification 关系指答案通过 notification 方式由传播者 $node_{src}$ 传播到接收者 $node_{dest}$, 且满足式(9),

$$\begin{aligned} & \forall rel \in E, rel[rtype] = notification \rightarrow \\ & (rel[node_{dest}][uid] \in Follower(Q)) \wedge \\ & rel[node_{src}][time] < rel[node_{dest}][time] \end{aligned} \quad (9)$$

qlink 关系指答案通过 qlink 方式由传播者 $node_{src}$ 传播到接收者 $node_{dest}$, 且满足式(10),

$$\begin{aligned} & \forall rel \in E, rel[rtype] = qlink \rightarrow \\ & \exists user \in U, \exists xact \in ACTION, (xact[uid] = user \wedge \\ & rel[node_{dest}][uid] \in Follower(user) \wedge \\ & ((user \in Follower(Q) \wedge xact[item_id] = qid) \vee \\ & (\exists A \neq A', xact[item_id] = aid(A') \wedge A' \in Ans(Q) \wedge \\ & user \in Upvoter(A') \cup Answerer(A')))) \wedge \\ & xact[time] < rel[node_{dest}][time] \end{aligned} \quad (10)$$

recommendation 关系指答案通过 recommendation 方式由传播者 $node_{src}$ 传播到接收者 $node_{dest}$, 且满足式(11):

$$\begin{aligned} & \forall rel \in E, rel[rtype] = recommendation \rightarrow \\ & rel[rtype] \neq follow \wedge rel[rtype] \neq notification \wedge \\ & rel[rtype] \neq qlink \end{aligned} \quad (11)$$

需要说明的是, 对于 Q 的每一个回答, 都存在一个定义 5 中的传播网络, 这些传播网络组成社会化问答网络知识传播模型。

定义 6. 社会化问答网站知识传播模型(M). 用二元组 $M = \langle Q, G_Q \rangle$ 描述, 其中 Q 为问题, G_Q 为社会化问答网站中一个问题 Q 的所有答案 A 的知识传播网络, 即 $G_Q = \{G_A\}$.

3.3.2 知识传播网络性质

从社会化问答网站的实际情况出发, 结合首次激活独立级联模型假设及知识传播网络的定义, 知识传播网络应满足如下性质, 这些性质将在传播网络推断过程中用到。

(1) 行为性质

- ① 回答节点只能作为传播者节点;
- ② 点赞节点可以作为传播者节点, 也可以作为接收者节点;
- ③ 评论、收藏节点只能作为接收者节点。

(2) 类型性质

- ① 传播者节点对应的用户至少拥有提问者、问题关注者、回答者、点赞者之一的身份;

② 除回答节点作为传播者外, 传播网上所有的传播者也是接收者。

(3) 度数性质. 答案从传播者传播到接收者, 每个入度大于零的节点对应一个接收者用户, 每个出度大于零的节点对应一个传播者用户。

(4) 时间性质. 传播起始节点时间小于传播终止节点时间。

(5) 对应性质. 每个节点和用户一一对应。

需要说明的是, notification 关系不需要除回答者之外的传播者, 所以由回答节点直接到达接收者。实际中, qlink 方式需要通过同一个问题下的其它回答的点赞者、回答者、或问题关注者传播, 本文为了简化模型, 以答案为单位定义传播模型, 不包含问题关注者和其它回答涉及的用户, 并认为 qlink 关系的起始节点是回答节点。

3.3.3 知识传播网络操作

定义 7. 添加接收者操作($AddReceiver$). 将一个新的接收者节点加入原有的知识传播网络, 并添加一条从已有节点到新节点的关系. 定义如式(12)所示:

$$G_{after} = AddReceiver(G, node_{old}, node_{new}, rtype) \quad (12)$$

G_{after} 表示操作后得到的传播网络, G 表示一个知识传播网络, $node_{old}$ 表示 G 中的一个节点, $node_{new}$ 表示将要添加的节点, $rtype$ 表示将要添加的关系。

4 传播网络推断方法

4.1 方法概述

4.1.1 问题定义

定义 8. 用户关注关系网络(F). 称 F 为答案传播网络 $G = (V, E)$ 对应的用户关注关系网络, 满足:

- (1) F 是一个有向图, 用 (V_F, E_F) 表示;
- (2) $V_F = V$;
- (3) $\forall rel \in E_F \rightarrow$

$$FO(G, rel[node_{src}], rel[node_{dest}]) = true;$$

- (4) $\forall node_1 \forall node_2 (node_1 \in V \wedge node_2 \in V \wedge FO(G, node_1, node_2) = true) \rightarrow$
 $\exists rel (rel \in E_F \wedge rel[node_{src}] = node_1 \wedge rel[node_{dest}] = node_2).$

在定义 8 中, 条件(3)指 F 中的每一条边都代表 G 中用户的关注关系. 条件(4)指 F 包含了所有可能的用户关注关系. FO 为节点关注关系判断函数, 见定义 9。

定义 9. 节点关注关系判断函数(FO). 称 $FO: \langle \langle G, node_{src}, node_{dest} \rangle \rangle \rightarrow \{true, false\}$ 表示在图 G 中 $node_{dest}[uid]$ 代表的用户是否关注了 $node_{src}[uid]$ 代

表的用户,若关注,返回 true,反之返回 false.

定义 10. 传播网络推断问题. 对于问题 Q 的一个答案 A , 已知 $L_{fq}, L_{up}, L_{cm}, L_{cl}$ 和用户关注关系网络 F , 推断答案 A 的传播网络 G_A . 即:

设一个问题 Q 的提问者的提问行为

$$Action_{ask} = \langle uid_q, Act_{ask}, t_q, qid \rangle,$$

回答行为 $Action_{answer} = \langle uid_a, Act_{answer}, t_a, aid \rangle$

已知: 按时间排序但具体时间未知的问题关注者关注问题行为列表 L_{fq} :

$$[\langle uid_{fq_1}, Act_{followQ}, null, qid \rangle, \langle uid_{fq_2}, Act_{followQ}, null, qid \rangle, \dots, \langle uid_{fq_m}, Act_{followQ}, null, qid \rangle]$$

$$uid_i \in Follower(Q), fq_1 \leq i \leq fq_m.$$

按时间排序但具体时间未知的点赞者点赞行为列表 L_{up} :

$$[\langle uid_{up_1}, Act_{upvote}, null, aid \rangle, \langle uid_{up_2}, Act_{upvote}, null, aid \rangle, \dots, \langle uid_{up_n}, Act_{upvote}, null, aid \rangle]$$

$$uid_i \in Upvoter(A), up_1 \leq i \leq up_n.$$

按时间排序的评论者评论行为列表 L_{cm} :

$$[\langle uid_{cm_1}, Act_{comment}, t_{cm_1}, aid \rangle, \langle uid_{cm_2}, Act_{comment}, t_{cm_2}, aid \rangle, \dots, \langle uid_{cm_o}, Act_{comment}, t_{cm_o}, aid \rangle]$$

$$uid_i \in Commenter(A), cm_1 \leq i \leq cm_o.$$

按时间排序的收藏用户列表 L_{cl} :

$$[\langle uid_{cl_1}, Act_{comment}, t_{cl_1}, aid \rangle, \langle uid_{cl_2}, Act_{comment}, t_{cl_2}, aid \rangle, \dots, \langle uid_{cl_p}, Act_{comment}, t_{cl_p}, aid \rangle]$$

$$uid_i \in Commenter(A), cl_1 \leq i \leq cl_p.$$

任意用户 u 的关注者集合 $Follower(u)$ 和被关注者集合 $Followee(u)$.

求解: 问题 Q 的一个答案 A 的传播网络 G_A .

定义 10 中关注问题和点赞行为的具体时间未知, 主要是由于问题关注者关注问题时间、点赞者点赞时间在行为发生的一段时间之后难以获取, 需要在传播网络推断过程中判断节点的时间先后次序, 对时间进行填充.

4.1.2 方法步骤

本文通过如下步骤进行传播网络推断.

步骤 1. 根据评论、收藏和回答时间, 对所有答案的 L_{up} 和 L_{fq} 进行时间最优填充.

步骤 2. 将已标注数据分成训练集和测试集两部分.

步骤 3. 对测试集生成接收者列表 L_r , 传播者列表 S , 用户-用户行为哈希表 $UserAction$.

步骤 4. 推断 follow 关系, 得到 follow 关系子图 G_f .

步骤 5. 从 L_r 中排除掉通过 follow 关系接收到答案的用户, 得到 L'_r .

步骤 6. 遍历 L'_r , 依次推断 notification 关系、qlink 关系.

步骤 7. 对未推断出传播方式的用户, 认为通过 recommendation 方式接收到答案.

步骤 8. 与 G_f 融合, 得到最终的传播网络 G .

步骤 1 中, “时间填充”指利用已知信息填充 L_{up} 中点赞行为和 L_{fq} 中关注问题行为的时间. 以填充 L_{fq} 为例, 我们可由此得知用户 u 关注问题 Q 和答案 A 被提交这 2 个行为的先后关系. 填充方法如下: 对每一个问题关注者, 利用用户行为列表找到该用户的回答、评论、收藏行为, 若存在这些行为, 则把其中最早的行为时间记为关注问题的时间. 由于对 L_{fq} 每个行为的时间填充是独立的, 很容易造成填充时间与问题关注者的相对顺序矛盾. 为避免矛盾同时获得尽可能多的时间信息, 本文将时间最优填充建模为最长上升子序列问题. 即对于一个长度为 n 的问题关注者序列 $[uid_{fq_1}, uid_{fq_2}, \dots, uid_{fq_n}]$, 已知其中若干关注者对应的时间, 构成一个序列 $T: [t_1, t_2, \dots, t_m]$, 求该序列的最长非递减子序列 $T': [t'_1, t'_2, \dots, t'_n]$, 满足 $t'_{i-1} \leq t'_i, i \in N \wedge i < LEN(T')$. 对于填充失败的时间, 使用 $TimeRange$ 表示, 具体见 4.2.3 节.

步骤 5 中, follow 关系推断转化为二分类问题, 使用 SVM 进行推断. 因为 follow 关系可能的候选边很多, 凡是一组 followee-follower 都是可能的传播者-接收者, 在这种情况下不可能通过每次观察一条边判断其是不是真的 follow 关系, 需要考虑候选边之间的联系. 比如候选传播者排序特征 (h_rank). 最后在选择 follow 关系边的时候, 一个接收者如果有多条候选边, 也是选择置信度最高的那一条, 这也是每次判断一条边不能做到的.

步骤 6 中, 根据定义, 采用启发式方法推断 notification, qlink 和 recommendation 关系. 因为这 3 种关系的传播者是已知的, 即回答者, 所以在判断某个接收者接收到答案的方式时, 只需推断传播方式即可. 针对回答节点到该接收者的边, 按照 $notification > qlink > recommendation$ 的优先级顺序, 首先推断是否为优先级最高的关系, 如果是则推断结束, 如果不是则判断是否为低一优先级的关系. 换句话说, 在推断后 3 种关系的过程中除了当前边之外并不需要考虑其它的边, 为降低算法复杂度, 采用启发式方法推断剩余 3 种关系.

4.2 follow 关系推断

4.2.1 基本思路

follow 关系推断的基本思想是在由用户关注关系构成的网络上求解 follow 关系边, 将此问题转化

为二分类问题,先根据用户关注关系网络生成所有可能的 follow 关系边,然后将这些边分为“是”和“不是”follow 关系 2 类。

定义 11. follow 关系子图(G_f). 称 G_f 为答案传播网络 G 的 follow 关系子图,当且仅当 G_f 是把 G 中所有 follow 关系的边收集起来构成的子图. follow 关系推断本质上就是要求出 G_f .

根据 G, G_f 和 F 之间的关系可以得到定理 1.

定理 1. 对于同一个答案传播网络 G, G_f 是用用户关注关系网络 F 的一个子图.

证明. 令 $G = (V, E), F = (V_F, E_F), G_f = (V_{G_f}, E_{G_f})$, 据子图的定义,我们只需要证明 $V_{G_f} \subseteq V_F$ 且 $E_{G_f} \subseteq E_F$ 即可.

由 G_f 是 G 的子图可知 $V_{G_f} \subseteq V$, 又因为 $V_F = V$, 故 $V_{G_f} \subseteq V_F$.

由 G_f 是 G 的子图可知:

$$\forall rel \in E_{G_f} \rightarrow rel[node_{src}] \in V \wedge rel[node_{dest}] \in V \quad (13)$$

由 G_f 定义条件 $E_{G_f} = \{rel | rel \in E \wedge rel[rtype] = follow\}$ 可知:

$$\forall rel \in E_{G_f} \rightarrow rel[rtype] = follow \quad (14)$$

由式(14)及 follow 关系的

$$\forall rel \in E_{G_f} \rightarrow$$

$$FO(G_f, rel[node_{src}], rel[node_{dest}]) = true \quad (15)$$

由式(13)、式(15)可知

$$\forall rel \in E_{G_f} \rightarrow rel[node_{src}] \in V \wedge rel[node_{dest}] \in V \wedge FO(G_f, rel[node_{src}], rel[node_{dest}]) = true \quad (16)$$

结合 F 定义条件(4)可知 $E_{G_f} \subseteq E_F$, 从而 G_f 是 F 的一个子图. 证毕.

定理 1 证明了 G_f 是 F 的一个子图,即对于 F 中的一条边,它要么是 G_f 中的一条边,要么不是. 目标是求出 G_f , 显然可以转化为一个二分类问题.

定义 12. 用户关注关系网络二分类问题. 对于答案传播网络 G 的用户关注关系网络 F , 把它的所有边按是否对应 G 中的一条 follow 关系边分为两类.

为解决此二分类问题,本文将数据集分为训练集和测试集 2 部分,按如下步骤推断 follow 关系:

步骤 1. 融合相同用户节点,生成候选边.

步骤 2. 从训练集提取特征,使用 SVM 模型训练出 follow 关系的分类器.

步骤 3. 在测试集上推断出所有 follow 关系,得到 follow 关系子图 G_f .

其中,步骤 1、步骤 2 的具体实施方法见 4.2.2 节和 4.2.3 节,步骤 3 中,若存在入度 > 1 的接收者节点,选择置信度最大的节点作为传播者节点. 如图 5 所示,假设用户 D 只关注了用户 B 和用户 C , “ $\rightarrow$ ”代表知识传播方向,用户 D 的入度为 2,若边 CD 和边 BD 都被标注成 follow 关系边,则选择其中可能性更大的那条作为推断结果.

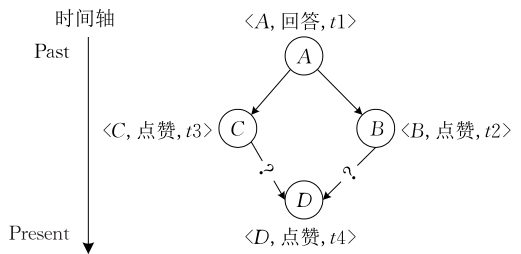


图 5 多人度节点推断示意图

4.2.2 候选边的生成方法

要解决用户关注关系网络二分类问题,需要不重不漏地生成候选边且处理一个用户多次出现的问题. 下面分 2 个步骤介绍候选边的生成方法.

(1) 融合相同用户节点

相同用户节点融合的目的是融合同一用户的多种操作,生成传播网络中的节点. 一个用户可能有多个操作,例如,用户“张三”可能同时存在于评论者列表和点赞者列表中. 若不对其进行处理,则候选边中将出现多个不同的节点,其对应的用户都是具有相同 uid 的“张三”.

融合结果用哈希表 H 存储,形如 $\{uid: node\}$. 图 6 展示了节点融合过程.

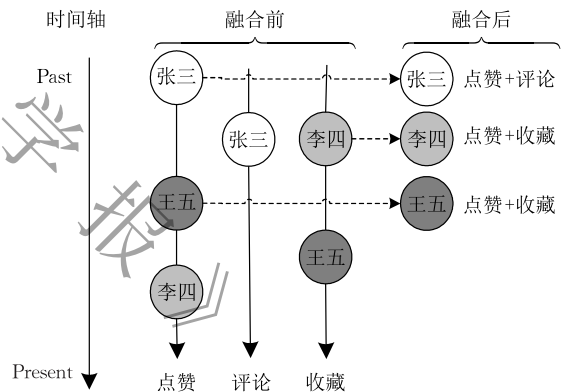


图 6 节点融合

(2) 生成候选边

候选边是指可能成为 follow 关系的边. 由数据结构定义可直接排除一些不合法边, 如由 follow 关系起始节点类型定义可知, 起始节点的类型为评论的边不可能对应 G 中的一条 follow 关系边. 在生成候选边的过程中, 若能排除掉尽可能多的边, 将减少数据量, 加快下一步训练模型的速度. 生成候选边的步骤如下:

步骤 1. 拼接回答者和点赞者列表 L_{up} , 构成 follow 关系候选起始节点列表 $L_{follow-sender}$.

步骤 2. 依次拼接点赞者列表 L_{up} 、评论者列表 L_{cm} 、收藏者列表 L_{cl} , 构成候选接收者列表 $L_{follow-receiver}$.

步骤 3. 使用两重循环分别遍历 $L_{follow-sender}$ 和 $L_{follow-receiver}$, 若当前遍历到的用户之间存在关注关

系, 则从 H 中获取融合后的节点, 构成候选边. 使用一个 $EdgeSet$ 记录已经生成的候选边, 防止重复.

使用这种方法, 排除了所有已知不可能是 follow 关系的边, 且没有遗漏.

4.2.3 特征提取

对于该二分类问题, 将从训练数据提取特征, 训练出 follow 关系的分类器. 使用已标注数据作为训练数据. 本文选取了 3 种特征, 由于分类的对象是用户关注关系网络 F 的边, 所以特征都是针对一条边而言.

(1) 节点类型特征

根据 follow 关系的定义, G_f 中的一条边起始节点类型的取值集合是 $\{Act_{answer}, Act_{upvote}\}$, 终止节点类型的取值集合是 $\{Act_{upvote}, Act_{comment}, Act_{collect}\}$ 的非空子集. 对关系 $rel = \langle node_{src}, node_{dest}, rtype \rangle$, 使用 3 个特征表示节点类型, 记作 $is_{upvote}, is_{comment}, is_{collect}$, 符合相应类型记为 1, 否则为 0.

(2) 起始节点和终止节点时间关系特征

由于起始节点的时间必然早于终止节点, 所以节点对应的用户操作的时间能够为推断提供信息. 鉴于传播网络推断中点赞者的操作时间无法确定, 有时无法确定起始节点和终止节点哪个时间更早, 在式(17)中用 uncertain 描述这种不确定. 因此用 3 个值表示时间关系特征, 记作 r_order :

$$r_order = \begin{cases} -1, & node_{src}[time] < node_{dest}[time] \\ 0, & \text{uncertain} \\ 1, & node_{src}[time] > node_{dest}[time] \end{cases} \quad (17)$$

为了尽可能获得准确的时间关系, 定义 $TimeRange(start_time, end_time)$, 表示从 $start_time$ 开始到 end_time 的时间范围. 在节点融合过程中, 已经根据评论时间和收藏时间确定了部分点赞者的时间. 接下来, 根据这些点赞者时间和点赞者的相对顺序可以确定其它点赞者的时间范围.

例如, 考虑下列长度为 n 的点赞者列表(为简洁省略了 Act_{upvote} 和 aid), 初始状态时间信息都是 null:

$$[\langle uid_{up_1}, null \rangle, \langle uid_{up_2}, null \rangle, \langle uid_{up_3}, null \rangle, \dots, \langle uid_{up_n}, null \rangle]$$

假设在节点融合过程中把 uid_{up_1} 的时间设置为 t , 即:

$$[\langle uid_{up_1}, null \rangle, \langle uid_{up_2}, t \rangle, \langle uid_{up_3}, null \rangle, \dots, \langle uid_{up_n}, null \rangle]$$

则可以用 $TimeRange$ 填充剩下的时间

$$[\langle uid_{up_1}, TimeRange(null, t) \rangle, \langle uid_{up_2}, t \rangle, \langle uid_{up_3}, TimeRange(t, null) \rangle, \dots, \langle uid_{up_n}, TimeRange(t, null) \rangle]$$

如果在节点融合过程中把 uid_{up_2} 的时间设置为 t , uid_{up_n} 的时间设置为 t' , 则填充后的时间是:

$$[\langle uid_{up_1}, TimeRange(null, t) \rangle, \langle uid_{up_2}, t \rangle, \langle uid_{up_3}, TimeRange(t, t') \rangle, \dots, \langle uid_{up_n}, t' \rangle]$$

在这个过程中, 如果出现时间范围矛盾的情况, 比如 $end_time < start_time$, 或者时间违背了相对顺序, 选择不填充时间.

在比较时间先后时, 若涉及需要比较 $TimeRange$, 按如下时间判断规则进行.

时间判断规则:

规则 1. 若 $node_{src}$ 和 $node_{dest}$ 类型都是点赞, 直接根据相对顺序比较.

规则 2. 若 $node_{src}$ 是回答节点, 则 $node_{src}$ 时间早于其它节点.

规则 3. 若 $node_{src}[time] = null$ 或 $node_{dest}[time] = null$, 则 $r_order = 0$.

规则 4. 若 $node_{src}[time] = TimeRange(s, e)$, $node_{dest}[time] = t$, 则

$$r_order = \begin{cases} -1, & t < s \\ 1, & e < t \\ 0, & \text{其他} \end{cases}$$

规则 5. 若 $node_{src}[time] = t$, $node_{dest}[time] = TimeRange(s, e)$, 则

$$r_order = \begin{cases} -1, & e < t \\ 1, & t < s \\ 0, & \text{其他} \end{cases}$$

规则 6. 若 $node_{src}[time] = TimeRange(s, e)$, $node_{dest}[time] = TimeRange(s', e')$, 则

$$r_order = \begin{cases} -1, & e < s' \\ 1, & e' < s \\ 0, & \text{其他} \end{cases}$$

(3) 候选传播者排序特征

候选传播者排序特征的含义为某个候选传播者在所有候选传播者中按时间的排名, 记作 h_rank . $h_rank = 0, 1, 2, \dots, N$, N 为自然数, 候选者的时间排名越靠前, 其 h_rank 越小. 如上文图 5 中, 用户 B 点赞的时间 t_2 早于用户 C 点赞的时间 t_3 , 因此边 BD 的候选传播者排序特征值为 0, 边 CD 的候选传播者排序特征值为 1.

4.3 其他关系推断

因为 follow 的优先级最高, 所以在推断 notification, qlink, recommendation 关系之前, 首先要将被标注为 follow 关系终止节点对应的接收者从接收列表 L_r 中去掉, 其中 $L_r = L_{up} + L_{cm} + L_{cl}$, 剩下的待推断的接收者, 记作 L'_r .

推断过程的总体流程是遍历接收者列表 L'_r , 按优先级假设从高到低依次判断, 如果能确定是通过某种传播方式接收到答案, 则不再继续, 否则继续判断是否可能是低优先级传播方式.

4.3.1 notification 关系推断

notification 方式推断规则: 若满足 2 个条件, 则认为属于问题 Q 的答案 A 通过 notification 方式

传播到用户 u ;

(1) $u \in \text{Follower}(Q)$

(2) 无法确定 u 关注 Q 的时间晚于答案 A 被提交的时间

首先判断 u 是否关注了 Q , 如果关注了, 再比较关注时间和答案 A 被提交的时间, 只要不能判断 u 关注 Q 的时间晚于答案 A 被提交的时间, 则认为 u 通过 notification 式接收到答案 A .

由于 L_{fq} 已进行了最优时间填充, 仿照提取起始节点和终止节点时间关系特征中的做法, 将剩余未填充的时间用 TimeRange 表示. 如果能够确定关注时间早于答案提交时间, 或者无法判断时间关系, 都认为 u 通过 notification 方式接收到答案 A . 这种做法在一定程度上损失了判断的精确性, 但通过 5.2 节传播网络推断实验可知, 在精确时间信息缺失的情况下仍能保证关系推断结果有较好的准确率.

4.3.2 qlink 关系推断

4.2.3 节在计算 follow 关系的起始节点和终止节点时间关系特征时已经对所有答案的点赞者列表 L_{up} 进行了时间填充. 为推断用户 u 是否通过 qlink 方式接收到答案 A , 本文定义了传播者列表 S 和用户-用户行为哈希表 UserAction . 候选传播者列表 S 收集所有可能的传播者对应的行为, 包括关注问题行为、点赞行为. 用户行为的 item_id 属性是该问题下任意答案的 ID 或问题 ID. 判断用户 u 是否通过 qlink 方式接收到答案, 只要遍历 S , 筛选出符合 qlink 关系候选传播者判断规则的用户, 就可以找到 qlink 关系的传播者. 用户-用户行为哈希表 UserAction 被用来记录所有用户的所有行为, 其结构形如 $\{key: value\}$, 其中 key 为用户 ID, $value$ 为该用户 u 在某问题下所有行为的列表

$$\text{action_list} = [\text{Action}_1, \text{Action}_2, \dots].$$

本文定义 ADD 函数, 将用户行为加入到 UserAction 中.

下面给出判断用户 u 是否通过 qlink 关系接收到答案 A 的推断算法, node_u 表示节点融合后用户 u 的节点:

算法 1. qlink 关系推断算法.

输入: $\text{aid}, L_{cm}, L_{cl}, L_{fq}, L_{up}, \text{node}_u$

输出: is_qlink

1. $S = \{\text{Act}_{ask}\} \cup L_{fq} \cup \{\text{Act}_{answer}\} \cup L_{up}$;
//构造候选传播者列表
2. $\text{SORT}(S)$ BY (ActionTime)
3. $\text{UserAction} = \emptyset$; //初始化用户-用户行为哈希表
4. $\text{ADD}(\text{uid}_q, \text{Act}_{ask})$ //记录提问
5. FOR each $\text{answer} \in A(Q)$ //遍历所有答案
6. $\text{ADD}(\text{uid}_a, \text{Act}_{answer})$ //记录回答
7. FOR each $\text{action} \in \{L_{cm} \cup L_{cl} \cup L_{up} \cup L_{fq}\}$
8. $\text{ADD}(\text{action}[\text{uid}], \text{action})$ //记录评论, 收藏, 点赞, 关注问题

9. END FOR
10. END FOR
11. FOR each action_list in UserAction ;
12. $\text{SORT}(\text{action_list})$ BY (ActionTime)
13. END FOR
14. FOR each action in $\text{UserAction}[u]$
15. IF $\text{action}[\text{time}] < \text{node}[\text{time}] \wedge \text{action}[\text{item_id}] \neq \text{aid}$
16. RETURN true //如果无法确定 u 和其它答案的交互操作晚于和答案 A 的交互操作, 则判定为 qlink
17. END IF
18. END FOR
19. FOR each action in S //遍历候选传播者列表
20. IF $\text{action}[\text{uid}] \in \text{Followee}(u) \wedge \text{action}[\text{time}] < \text{node}_u[\text{time}] \wedge \text{action}[\text{item_id}] \neq \text{aid}$
21. RETURN true //找到了 u 的 followee u' , u' 和其它答案进行过交互, 且不能确定 u' 的行为时间晚于 u 和答案 A 交互的时间
22. END IF
23. END FOR

算法 1 的第 1 行构造了候选传播者列表, 首先加入提问和关注问题行为, 然后遍历所有答案, 加入回答和点赞行为, 最后排序. 算法的 3~10 行构造了用户-用户行为哈希表, 将用户 ID 与用户行为对应, 并对每个用户的 action_list 进行排序.

如果用户在 t 时刻成为了另一个答案的接收者, 在 $t + \Delta t$ 时刻成为了答案 A 的接收者, 那么根据定义, 用户是通过 qlink 方式接收到答案 A 的. $\text{UserAction}[u]$ 记录了用户 u 在该问题下的所有操作, 14~18 行中, 遍历 $\text{UserAction}[u]$, 如果 u 在和答案 A 的交互操作之前, 可能有和其它答案的交互操作(回答、点赞、评论、收藏), 即无法确定和其它答案的交互操作晚于和答案 A 的交互操作, 则认为用户 u 通过 qlink 方式接收到答案.

19 至 23 行, 遍历传播者集合 S , 如果找到了 u 的 followee u' , 且不能确定 u' 的行为时间晚于 u 和答案 A 交互的时间, 就认为答案由 u' 通过 qlink 方式传播到 u . 最后返回结果.

4.3.3 recommendation 关系推断

根据定义, 如果无法确定一个用户是通过 follow, notification, qlink 关系接收到答案的, 就认为是通过 recommendation 方式. 根据 recommendation 关系起始节点约束, 在传播网络 G 中, recommendation 关系的起始节点是回答节点.

5 实验与效果评估

5.1 实验概述

5.1.1 实验数据

本实验采用第三方开源库 `zhihu-py3` (<https://github.com/7sDream/zhihu-py3>), 通过约 2 周时间

(2016. 2. 20~2016. 3. 6)采集了知乎上 4 个话题的问题和答案,这 4 个话题分别为互联网、生活、第三产业(服务业)和自然科学.在兼顾热门问题的同时具有一定的覆盖性,使得实验结果具有普遍性.之后采用文献[18]中的方法对采集的数据进行自动标注.

过滤没有回答的问题后,获得了 8344 个问题和对应的 14 227 个答案,以及 77 258 个相关用户的 47 446 个用户行为(点赞行为 24 954 个、评论行为 7408 个、收藏行为 857 个、其他为回答行为)和 62 000 个用户关注关系,在后续实验中,把其中的 7174 个问题和对应的 11 837 个答案用于训练,剩下的 1170 个问题和对应的 3440 个答案用于测试.

本文统计了答案数与问题数的对应关系,即答案数为 $N(N=1,2,3\cdots)$ 的问题数有多少个,结果如图 7 所示.数据分布基本符合长尾效应,答案数为 1 的问题有 3926 个,答案数为 2 的问题有 1195 个,随着答案数的增加,问题数迅速递减,当答案数为 21 及以上时,基本只有 1 个问题.



图 7 答案数-问题数对应关系图

5.1.2 评价指标

由传播模型的定义可知,推断出的传播网络和真实的传播网络相比,节点集合 V 完全相同,差异在于关系集合 E (分别为 follow, notification, qlink, recommendation 关系).因此,传播网络推断的准确程度实际上可由关系推断的准确程度衡量.定义 3 个评价指标:

(1)关系类型准确度 RTA (Relation Type Accuracy)

$$RTA = \frac{T_{\text{rtype}}}{N} \times 100\%$$

(2)传播者准确度 SA (Sender Accuracy)

$$SA = \frac{T_{\text{src}}}{N} \times 100\%$$

(3)关系准确度 RA (Relation Accuracy)

$$RA = \frac{T_{\text{rel}}}{N} \times 100\%$$

其中, T_{rtype} 代表关系类型正确推断的关系数量, T_{src} 代表传播者正确推断的关系数量, T_{rel} 代表关系类型正确推断且传播者正确推断的关系数量, N 代表总关系数量.

5.2 实验结果

5.2.1 follow 关系推断结果

(1)特征有效性分析

使用 SVM 分类器, k -fold 交叉验证与均方差 (MSE) 方法衡量特征有效性, 计算采用所有特征和每次排除掉一个特征时的 MSE 值, 结果如图 8 所示, 排除任意一个特征之后 MSE 值均会变大, 训练结果变差. 故本文选取 $h_rank, is_upvote, is_comment, is_collect, r_order$ 特征作为有效特征进行后续实验.

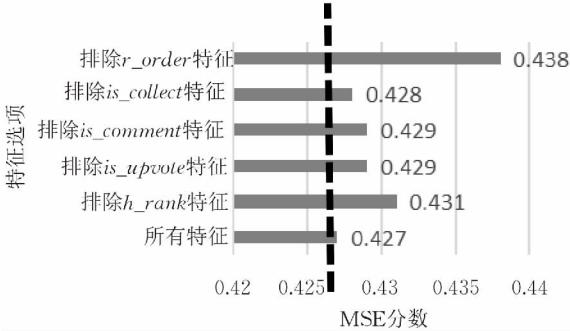


图 8 特征选择结果

(2) follow 关系分类结果

采用网格搜索法,在多种模型参数组合中进行选取,得到最优模型参数组合 $\{ 'C': 1000, 'kernel': 'rbf', 'gamma': 0.001 \}$, 在训练集上进行训练. 使用训练得到的分类器在测试集上进行测试, 结果如表 1 混淆矩阵所示.

表 1 测试集 follow 关系推断结果混淆矩阵

	推断为 0	推断为 1
实际为 0	$TN=1412$	$FP=1281$
实际为 1	$FN=1205$	$TP=4675$

该混淆矩阵表示直接将分类器应用于测试集进行 follow 关系标注的结果, 0 表示非 follow 关系, 1 表示 follow 关系.

分类精确度为 0.71, follow 关系准确率为 0.785, 召回率为 0.795. 需要特别说明的是, 以上结果并非最终的 follow 关系判断结果, 而是直接应用分类器得到的结果.

5.2.2 传播网络推断结果

得到传播网络推断结果后, 排除掉接收者数量为 0 的答案, 对每个答案分别计算 SA, RTA 和 RA , 绘制如图 9 所示散点图. 可以看到, 大多数答案的 SA, RTA 和 RA 在 0.95 以上, 多个点聚集形成一条黑色直线. 但存在一些答案的 RTA, RA 准确度非常低, 而 SA 是 1. 这是因为这些答案只有很少量的接收者, 传播者一般就是回答者, 所以 SA 很高, 而一旦关系类型推断错误, 则很可能出现 RA 和 RTA 是 0 的情况.

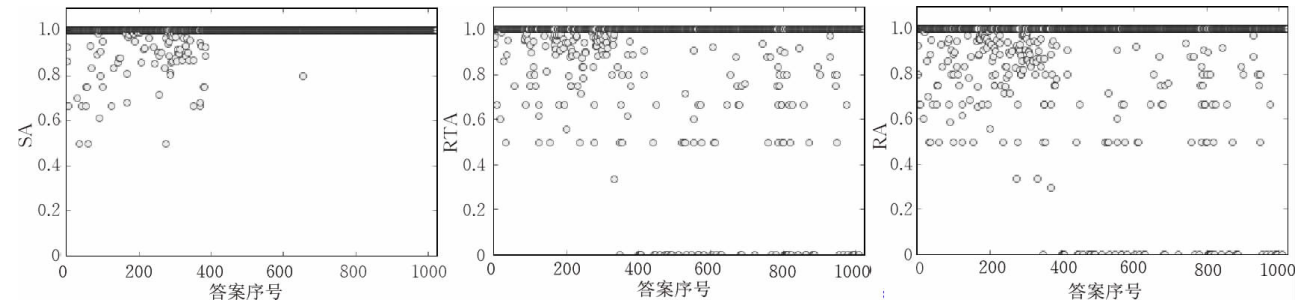


图 9 每个答案的传播网络推断结果准确度

进一步地,分别对 follow 关系推断结果、notification 关系推断结果、qlink 关系推断结果、recommendation 关系推断结果、所有答案的传播网络推断结果以及每个答案的传播网络推断结果的平均值进行了准确率分析,根据 5.1.2 节的评价指标计算方法,得到如图 10 所示结果.

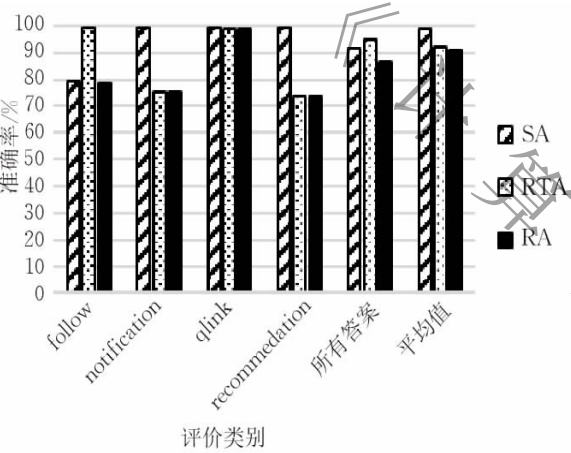


图 10 所有答案的传播网络推断结果准确度

图 10 中,4 种传播方式的 SA,RTA 和 RA 均在 73%以上,最高可达 99%. 由于前文定义 notification, qlink 和 recommendation 方式的传播者都是回答节点,只要没有将其推断成 follow 方式,其传播者一定正确,故其 SA 高达 99%,而 RA 和 RTA 存在低于 80%的情况. 在所有答案的传播网络中,SA 到达 0.948,RTA 达到 0.914,RA 达到 0.867,可见传播网络推断有令人满意的结果. 对每个答案的 SA, RTA 和 RA 取平均后平均准确度仍然很高,这说明大部分答案的传播网络推断准确度很高.

5.2.3 用户行为对知识传播的影响

本节通过行为数据统计,初步分析了用户行为对知识传播广度(接收者数量)和深度(传播距离)的影响,点赞、收藏和评论行为影响知识传播,点赞数和传播距离正相关.

(1) 用户行为对答案接收者数量的影响

为了验证本文中定义的合理性,将知识接收者数量与点赞数、收藏数和评论数进行对比. 结果如图 11 所示.

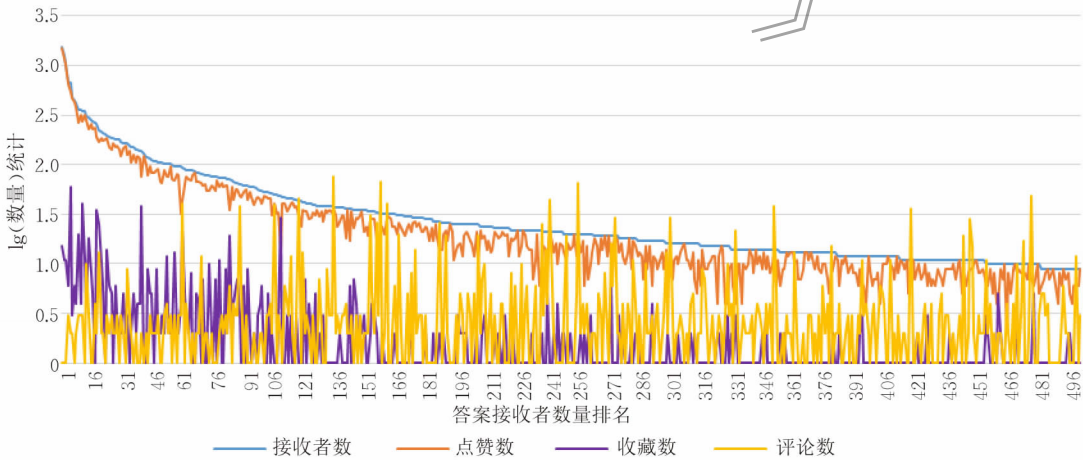


图 11 接收者相关数据对比

本文定义的答案接收者为某答案的点赞者、评论者和收藏者. 为了使示意图更直观,本文随机选取了近 500 个答案,统计其点赞数、收藏数和评论数,由实验数据可知,收藏数和评论数相对较少,分布较

均匀. 点赞数基本与接收者数量呈正相关,而评论数与收藏数出现波动. 点赞、收藏、评论 3 种用户行为影响知识传播,点赞数是决定答案接收者数量的主要因素.

(2) 用户行为对传播路径长度的影响

本文统计了每个答案的传播深度,结果如表 2 所示.

表 2 每个答案传播深度统计

答案传播深度	答案数
1	3376
2	273
3	76
4	27
5	16
6	7
7	3
8	3
9	1

从表 2 中可以看到传播深度为 1 的答案高达 3376 个,深度为 2 的答案为 273 个,大多数答案的传播深度在 9 以内,由传播网络定义可知,点赞行为是可以传播的,故本文分析了答案传播深度与点赞行为的关系,如图 12 所示.

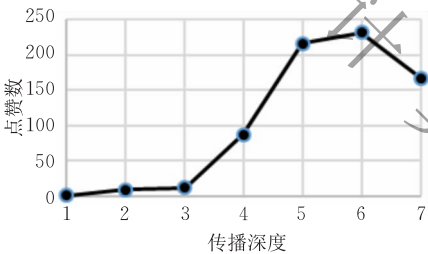


图 12 答案传播深度与用户行为的关系

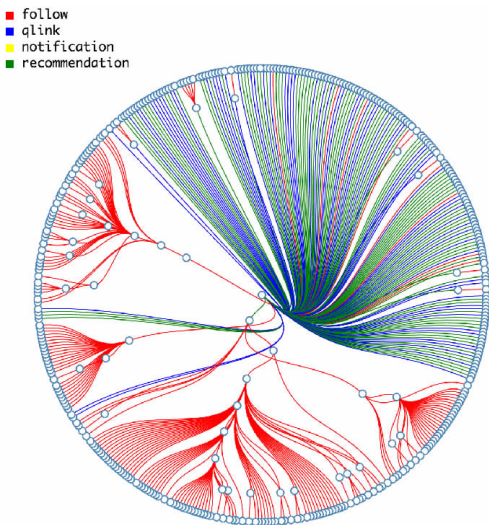
当传播路径深度在 6 以内时,深度随点赞数量的增加而增加,当深度大于 6 时点赞数有所下降,说明传播深度还受其他因素影响,比如用户关注关系等,在 5.3 节的实例分析中将进一步阐述.

5.3 实例分析

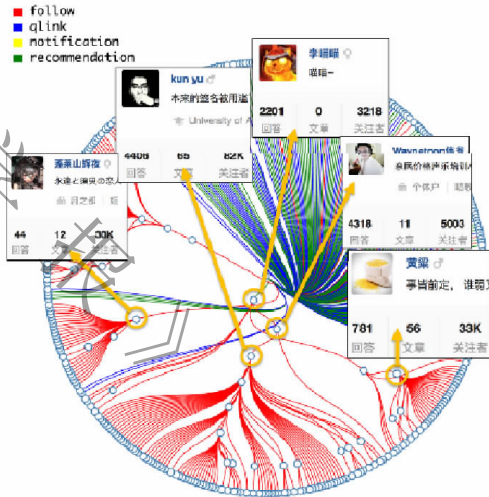
本节将以前一段时间的热门话题“上海女孩逃离江西”作为实例,展示传播网络推断的结果,并对传播过程以及造成大面积传播的原因做简要分析.该问题及答案截图如图 13 所示(问题链接:https://

www.zhihu.com/question/40598855/, 答案链接: https://www.zhihu.com/question/40598855/answer/87343917).

本文使用 d3.js 对传播网络进行可视化,该答案的传播网络图如图 14 所示.



(a) 答案传播网络可视化结果



(b) 关键节点示意图

图 14 答案传播网络可视化结果及关键节点示意图

图 13 和图 14(a)中接收者数目不同,因为 2 张图反映的并不是同一个时刻的情况.通过观察图 14,可以得到一些结论:

- (1) 答案的最大传播距离为 9;
- (2) 传播网络中最多的是 follow 关系,其次是 qlink 和 recommendation 关系;
- (3) 传播过程中有若干关键节点(图 14(b)中圆圈标注的点),它们的存在带来了更多的传播者.

下面图 14(b)中下方的关键节点对应的用户为例进行分析.此部分知识的传播路径如图 15 所示.



图 13 知乎某问题及答案截图

回答者 → 李喵喵(3218 followers) → 蓬莱山辉夜(30k followers)
 ↳Waynetong 伟爷(5003 followers) → 黄梁(33k followers)
 ↳Kun yu(82k followers)

图 15 传播路径展示

答案通过两个中等 follower 数量用户“李喵喵”和“Waynetong 伟爷”传播到若干拥有上万 follower 的用户“蓬莱山辉夜”、“kun yu”、“黄梁”，形成大面积传播。

通过对该传播过程实例中的关键节点的分析可以看出，一个答案要传播得更深更广，吸引到拥有大量 follower 的用户（以下称大 V 用户）是非常重要的。然而，对于普通用户而言，想直接成为大 V 用户的 followee 是很难的，因此我们看到，一个答案的扩散往往不能直接传播到大 V 用户，而是通过一些有数量适中的 follower 的用户间接地传播到大 V 用户。

图 16 展示了该问题下的若干其它答案的知识传播网络可视化结果，可以形象地看出不同答案传播的深度和广度。

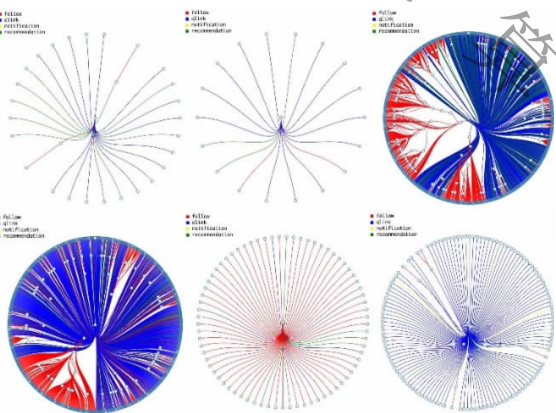


图 16 某问题下若干其它答案传播网络可视化结果

通过案例分析，不同知识的传播广度与深度不同，这与知识传播者关注者数量、用户行为等原因密不可分。

6 结束语

本文提出了一种描述社会化问答网站知识传播过程的传播网络模型，定义了数据结构和模型操作，为传播网络推断奠定了理论基础。提出了一种社会化问答网站知识传播网络推断方法，通过提取节点类型、节点时间关系及候选传播者排序特征推断 follow 关系，采用启发式方法推断另外 3 种关系，生成知识传播网络。实验结果表明推断方法有效，点赞行为和用户关注行为对知识传播的深度和广度具有一定的影响。

未来研究工作将基于本文方法推断出的知识传

播网络，深度挖掘社会化问答网站信息传播特点，进一步分析信息扩散规律及影响，预测传播趋势，逐步应用于网络舆情、专家发现、精准营销、推荐系统等领域。

参 考 文 献

- [1] Li Dan. A comparative study of Chinese and American online Q&A community—A case study of Quora and knowledge. Youth Journalist, 2014, (26): 19-20(in Chinese)
(李丹. 中美网络问答社区的对比研究——以 Quora 和知乎为例. 青年记者, 2014, (26): 19-20)
- [2] Li B, Jin T, Lyu M R. Analyzing and predicting question quality in community question answering services//Proceedings of the 21st International Conference Companion on World Wide Web. Lyon, France, 2012: 775-782
- [3] Zhou Xiaoqiang, et al. Answer sequence learning with neural networks for answer selection in community question answering. arXiv preprint arXiv:1506.06490, 2015
- [4] Liu Y, Bian J, Agichtein E. Predicting information seeker satisfaction in community question answering//Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Singapore, 2008: 483-490
- [5] Shah C. Measuring effectiveness and user satisfaction in Yahoo!Answers. First Monday, 2011, 16(2): 2-7
- [6] Riahi F, Zolaktaf Z, Shafiei M, et al. Finding expert users in community question answering//Proceedings of the 21st International Conference Companion on World Wide Web. Lyon, France, 2012: 791-798
- [7] Liu X, Ye S, Li X. ZhihuRank: A topic-sensitive expert finding algorithm in community question answering websites //Proceedings of the 14th International Conference on Web-Based Learning. Guangzhou, China, 2015: 165-173
- [8] Zhou T C, Lyu M R, King I. A classification-based approach to question routing in community question answering//Proceedings of the 21st International Conference Companion on World Wide Web. Lyon, France, 2012: 783-790
- [9] Wang Gang, et al. Wisdom in the social crowd: An analysis of quora//Proceedings of the 22nd International Conference on World Wide Web. Rio de Janeiro, Brazil, 2013: 1341-1352
- [10] Rodrigues E M, Natasa M. Socializing or knowledge sharing?: characterizing social intent in community question answering//Proceedings of the 18th ACM Conference on Information and Knowledge Management. Hong Kong, China, 2009: 1127-1136
- [11] Lu Lin-Yuan. Link prediction on complex networks. Journal of University of Electronic Science and Technology of China, 2010, 39(5): 651-661(in Chinese)
(吕琳媛. 复杂网络链路预测. 电子科技大学学报, 2010, 39(5): 651-661)
- [12] Zhou Tao, Lv Linyuan, Zhang Yicheng. Predicting Missing links via local information. The European Physical Journal B, 2009, 71(4): 623-630

- [13] Ou Q, Jin Y D, Zhou T. Power-law strength-degree correlation from resource-allocation dynamics on weighted networks. *Physical Review E*, 2007, 75(2): 021102
- [14] Clauset A, Moore C, Newman M E J. Hierarchical structure and the prediction of missing links in networks. *Nature*, 2008, 453(7191): 98-101
- [15] Kempe D, Jon K, Éva T. Maximizing the spread of influence through a social network//*Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Washington, USA, 2003: 137-146
- [16] Rogers E M. *Diffusion of Innovations*. New York: Simon and Schuster, 2010
- [17] Zafarani R, Mohammad A A, Huan L. *Social Media mining: An Introduction*. Cambridge: Cambridge University Press, 2014
- [18] Zuo Yao. *Analysis and Application of Knowledge Diffusion Network Inference in Community Question Answering Sites* [M. S. dissertation]. Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 2016(in Chinese)
(左遥. 社会化问答网站知识传播网络推断方法研究及分析应用[硕士学位论文]. 中国科学院计算技术研究所, 北京, 2016)



ZUO Yao, born in 1991, M. S. His research interests include big data analysis and mining.

LIANG Ying, born in 1962, M. S., senior engineer. Her research interests include data mining, big data process, ubiquitous computing etc.

BI Xiao-Di, born in 1992, M. S. candidate. Her research interests include big data analysis and mining.

SHI Hong-Zhou, born in 1971, Ph. D., senior engineer. His research interests include embedded system, ubiquitous computing, Internet of Things security.

DONG Xiang-Xiang, born in 1996, M. S. candidate. His research interests include big data analysis and mining.

LI Jin-Tao, born in 1962, Ph. D., professor. His main research interests include social media computing, multimedia content analysis, network computing technology etc.

Background

Information diffusion model and knowledge diffusion process have many applications in viral marketing, which is very important in current environment thanks to the emergence of Community Question Answering websites (CQA). It combines the advantages of social networking websites, Wikipedia and traditional Q&A websites. Current research of CQA websites is limited to analyzing raising and answering questions, such as finding experts for certain questions, predicting questioner satisfaction and evaluating answer quality, etc. At the same time, current research for CQA websites lacks attention for some hot research topics in social networking sites, such as how friend relationship influence information diffusion, how new messages spread among the network, and predicting hotness of new messages. The main reason is that social networking site records every sender of information propagation, while CQA website only records the overall situation, thus diffusion process data cannot be obtained from CQA website.

This paper aims at the key technologies relating to CQA websites knowledge diffusion process, including knowledge diffusion model in CQA website, knowledge diffusion process inference and knowledge diffusion process analysis system. We define the knowledge diffusion process from a sender to a receiver in CQA, and propose a diffusion network model to describe it. We also sum up four knowledge transmission modes, including follow, notification, qlink and recommendation. Based on the model, we propose a method to infer knowledge diffusion network. We use Zhihu as an example to verify the effectiveness of our methods.

This work is supported by the National Key Research and Development Program of China (2016YFB0800403), the National High Technology Research and Development Program (863 Program) of China (2015AA015803) and the Science and Technology Project of Beijing (Z161100001616009). It provides theoretical basis for big data analysis and mining.