StackExchange 问答社区知识传播①

陈风雷, 傅晨波, 宣 琦

(浙江工业大学 信息工程学院, 杭州 310023)

摘 要: StackExchange 是目前最流行的问答社区集结地之一. 本文利用 StackExchange 中具有美国地理信息的用户构建 StackExchange 问答社区在美国境内的知识传播图谱, 对传播网络的统计特征进行了分析, 提取出问答社区类网站的传播模式, 获取得到网络用户的知识分享方式. 我们发现 StackExchange 中的问答社区在分享知识过程中, 传播源往往不止一个. 同时, 我们为问答社区构建了知识传播图谱, 发现这些传播图谱具有相似的统计特征, 这意味着不同的问答社区可能具有类似的知识传播模式.

关键词: StackExchange; 地理信息; 问答社区; 演化; 传播网络

Knowledge Spreading in StackExchange Q&A Community

CHEN Feng-Lei, FU Chen-Bo, XUAN Qi

(College of Information Engineering, Zhejiang University of Technology, Hangzhou 310032, China)

Abstract: StackExchange is one of the most popular Q&A rendezvous, including dozens of Q&A communities. In the paper, based on the Q&A relationships and the geographical information of users, this paper analyzes the knowledge spreading patterns in Q&A communities in StackExchange in USA and statistical characters of networks. It extracts spreading patterns of Q&A communities and obtains knowledge sharing methods of users. We find that, generally, a Q&A community may have several spreading sources. Meanwhile, it establishes a spreading network for each Q&A community and finds that most of these spreading networks share similar statistical characters, indicating that these Q&A communities have similar spreading patterns.

Key words: StackExchange; geographical information; Q&A community; evolution; spreading

随着信息产业的发展以及互联网的迅速普及, 基于互联网的知识分享体系逐渐形成, 各类问答网站应运而生[1-3]. StackExchange[4]是一系列非常流行的问答社区的集合, 包含编程、物理、数学、科技文化及游戏等不同领域. 其中, StackOverflow [5]作为 StackExchange 中最著名的面向程序设计的问答网站之一, 得到了国内外很多研究人员的关注. 研究者们通过对 StackOverflow 各种特性的研究, 如用户名誉值[6-8], 不同年龄阶段的用户的贡献程度[9], 帖子[10-12], 标签[13]以及男女比例[14]等, 来了解 StackOverflow 的结构, 这些研究为研究者们研究 StackExchange 中其它社区提供了诸多宝贵的经验.

近年来, 对 StackExchange 的研究逐渐由单问答社区过渡到多问答社区. Posnett 等人[15]在研究 StackExchange 的问答社区时发现, 随着社区的发展, 由于追求名誉值的非专家用户参与度大于专家级用户或者专家级用户的名誉值得分达到了瓶颈, 导致答案的整体质量普遍降低. MacLeod 等人[16]通过构建问答社区标签网络, 发现用户的名誉值跟用户贡献多样性的标签相关. Furtado 等人[17] 通过分析不同问答社区贡献者的行为特征, 发现提供高质量贡献的用户并不是很积极, 且用户活动和专家的资料对于任务分配机制、专家识别以及社区管理的发展有帮助. 此外, 许多研究人员还对普通用户的行为进行研究. 如 Lal 等人

① 基金项目:国家自然科学基金(61273212,61572439,11505153);浙江省自然科学基金(LQ15A050002)
收稿时间:2016-01-19;收到修改稿时间:2016-03-14 [doi:10.15888/j.cnki.csa.005386]

[18]首次对 StackExchange 中的转移问题进行研究, 通过识别转移问题的特征, 提出机器学习的框架, 该模型在预测转移问题的准确率最大可达到 73%. Thongtanunam 等人[19]通过分析用户过去获得的奖励来找到用户的专业技能. Schenk 等人[20]通过将个人知识交易整合到地理信息水平的方式, 评估 StackOverflow 问答社区中知识经济的状态. 该研究考虑了有地理信息用户的答案被提问者采纳的情况, 并把这种采纳关系和相应的得分关系映射到国家水平, 进而研究国家间问答采纳关系以及财富值的分布状况. 最后作者探讨了随着时间的推移, 用户在世界各国的变化. 虽然 Schenk 等人对国家间问答采纳关系的分布状况以及用户在世界范围内的分布进行了探讨, 但并没有对国家内不同省份或州间的问答情况进行分析. 基于此, 本文在 Schenk 等人开创性工作的基础上分别探讨了 StackExchange 中的多个问答社区在美国各州之间的知识传播模式.

本文分析了 StackExchange 中 23 个问答社区数据, 研究了美国境内的知识分享传播模式, 并对知识传播网络的统计特征进行了分析, 从中提取出问答社区类网站的传播模式. 本文的余下部分结构如下: 第一节介绍了 StackExchange 网站的 23 个问答社区的数据集; 第二节详细描述 23 个问答社区的美国各州的知识传播图谱的构建方法; 第三节给出传播图的统计特征; 第四节对传播图以及统计特征进行分析; 最后一节给出总结和讨论.

1 问答社区数据的概览

1.1 数据简介

我们采集了 StackExchange 网站的 23 个问答社区从成立到 2014 年 5 月 4 日之间的数据, 每个问答社区包含如下 7 种主要类型数据, 分别为徽章(Badges)、帖子(Posts)、帖子历史(PostHistory)、帖子链接(PostLinks)、评论(Comments)、用户(Users)、投票(Votes). 我们主要考虑帖子数据和用户数据. 帖子数据包含帖子 Id、帖子类型 Id、帖子创建日期、帖子拥有者 Id(用户 Id 的一种)、父辈 Id(用户 Id 的一种, 帖子类型为答案时存在). 而用户数据包含用户 Id 和用户地址.

1.2 问答关系网络的可视化

在实际的数据中, 用户数据与帖子数据是分开存储的, 两者之间没有直接联系. 帖子中的问题帖子和答案帖子是根据时间关系排序的, 因此无法直接从数据中得到问题帖子和答案帖子之间的问答关系. 为了构建问答关系网我们首先要找到帖子的问答关系, 然后构建用户间的问答关系.

利用帖子与用户的信息, 我们可以构建出用户与用户的问答关系网. 具体来说, 根据帖子的类型, 我们可以将帖子分成问题帖子与答案帖子, 同时根据帖子对应的用户 Id, 也可以将用户分成问题用户及答案用户. 通过帖子之间的问答关系, 我们就可以构建用户之间的问答关系网络. 需要指出, 在用户问答中, 我们不考虑问题用户 Id 和答案用户 Id 相同的情况, 即自问自答的情况.

在本文中我们主要关注具有地理信息的用户. 将具有相同问答关系的用户进行求和作为用户间的问答权重, 我们可以得到美国与世界有地理信息用户的问答关系. 图 1 和图 2 分别给出了美国与世界用户的问答关系示意图, 其中节点代表用户, 连接代表问答关系, 深浅代表问答权重. 对比两图, 我们发现虽然 StackExchange 发源于美国, 但经过一段时间演化之后, 世界用户对 StackExchange 的贡献已经不可忽略了.

1.3 美国数据与世界数据

在图 3 中我们分别统计了 StackExchange 的 23 个问答社区中美国的用户数、帖子数以及用户问答总数占全世界的比值, 其中横坐标代表不同问答社区, 纵坐标代表美国用户数据与世界用户数据的比值. 从图中我们可以发现, 与其它问答社区相比, MathOverflow 作为面向专业数学家的问答社区, 其比值比较小, 这暗示着美国用户相比其它问答社区的用户, 在面向专业数学的领域, 美国用户的参与度并不高, 用户间的互动也相对较低. 而 Science Fiction&Fantasy 社区则恰恰相反, 作为面向科幻爱好者的问答社区, 美国用户无论是用户占比, 还是发帖数量比值都较大, 意味着美国用户对科幻类的事物较感兴趣, 热衷且积极参与, 互动频繁, 这与美国作为科幻电影高产国及其发达的科技密切相关. 由于 StackExchange 发源于美国, 本文将主要研究 StackExchange 中 23 个问答社区在美国境内的知识传播图谱.

2 研究方法

为了研究 StackExchange 中不同问答社区在美国

的知识传播图谱，我们将美国用户与用户之间的问答关系映射到美国州与州之间的问答关系，从而构建基于美国州的知识传播图谱，确定传播的特征时间.

2.1 美国州问答关系构建

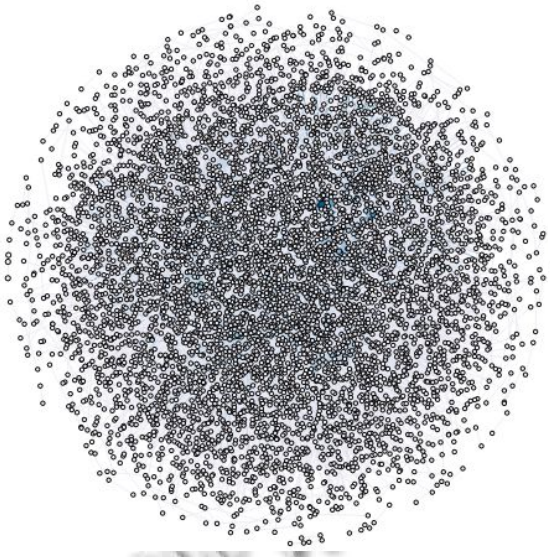


图 1 有地理信息的世界用户的问答关系示意图

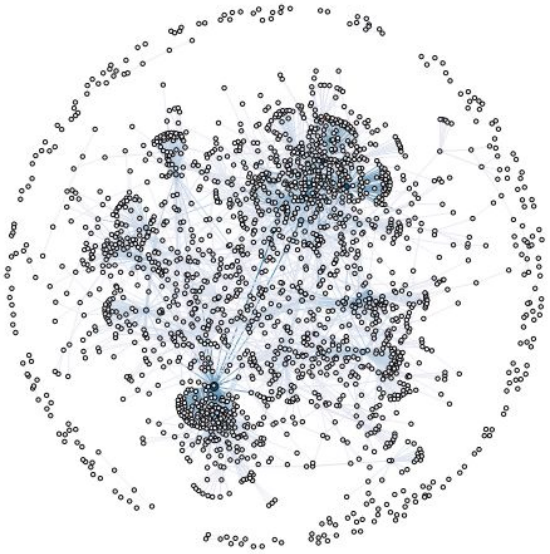


图 2 有地理信息的美国用户的问答关系示意图

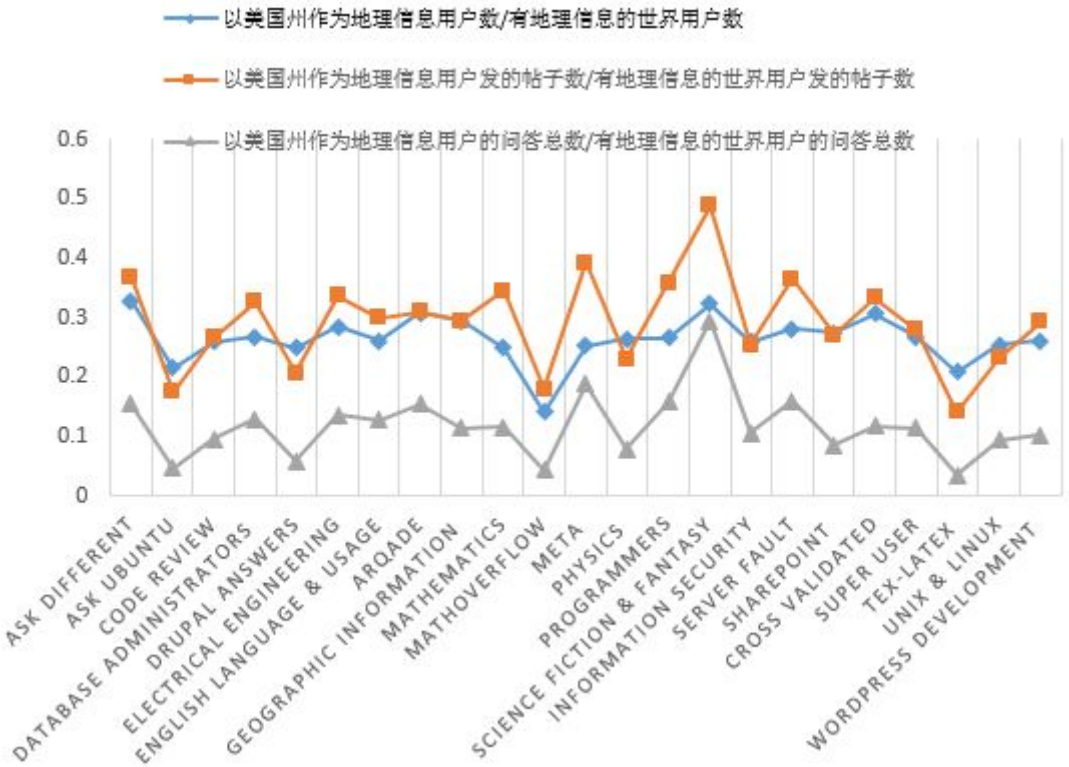


图 3 美国用户数据与世界用户数据的比值分布图

本文设计了以下算法，从帖子的问答关系中提取美国州的问答关系，具体步骤如下:

1) 在数据集中找出所有标有美国州地理信息的用户.

2) 统计出以美国州作为地理信息的用户发表的所有帖子，并根据帖子类型 Id 提取出所有的问题帖子和答案帖子.

3) 在问题帖子和答案帖子中，统计用户发表问题帖子的时间和发表答案帖子的时间，根据问题帖子的

拥有者 Id 以及答案帖子的父辈 Id,找出所有具有问答关系的问题帖子和答案帖子以及两者发表的时间.

4) 根据用户的地理信息, 将用户发表帖子和时间映射到美国各州发表帖子和时间上, 得到提出问题帖子的州(问题州)和回答问题帖子的州(答案州)的问答关系以及问题州发表帖子的时间(问题州时间)和答案州作答的时间(答案州时间).

2.2 特征时间 ΔT_i 的确定

在将帖子的问答关系映射到州与州之间的问答关系之后, 我们接下来就要确定传播的特征时间 ΔT . ΔT 的选取对研究美国州的传播至关重要, 太长或太短都不能反映知识传播的特征. 在本文中我们定义美国州 i 第一次出现的时间记为 T_{1i} , 州 i 第一次与任一其它州发生关系的时间记为 T_{2i} . 我们把 T_{1i} 和 T_{2i} 之间的时间间隔作为传播的特征时间 ΔT_i, 在 ΔT_i 范围内州 i 与先前出现州发生问答关系则代表先前州的影响已经传播到了该州. 如果在确定时间 T_{2i} 的过程中出现 $T_{1i}=T_{2i}$ 的情况, 我们选取与州 i 第二次发生关系时的答案州时间作为新的时间 T_{2i} , 其它情况以此类推.

值得注意的是, 在此定义下每个州的特征时间都是不同的, 这么做的好处是特征时间 ΔT_i 不会太长也不会太短, 同时在 ΔT_i 时间范围内也有极大可能出现与先前出现的州有互动行为. 在特征时间 ΔT_i 内, 确定前面出现的所有州对后面出现州的影响, 我们就可以构建出 23 个问答网站的知识传播关系图. 在图 4 中我们展现了部分问答社区的传播图, 其中节点代表美国州, 节点的大小和颜色深浅表示该州与其它州的互动情况, 节点越大, 颜色越深, 表示该州更热衷参与不同州的问答活动. 箭头方向代表传播方向, 连结的粗细表示州与州之间的互动程度, 连边越粗, 表示两州之间互动越频繁.

3 传播图的统计特征

值得指出的是, 不同问答社区的传播特性有可能是不同的. 为了显示不同问答社区之间的区别, 我们统计了知识传播网络的一些特征. 这些特征可以帮助我们更好地理解问答社区的知识传播特性. 在本小节, 我们主要统计了平均度、入度异质性、出度异质性、平均聚类系数、平均度、平均路径长度以及网络直径.

3.1 平均度

网络的度表示一个节点与网络中其它节点的连接情况. 在知识传播网络中, 度 k_i 表示与州 i 连接的所有其它州数目, 即州 i 的邻居数. 度越大, 说明该州的问答关系越广, 传播范围也越广. 传播网络的平均度是所有度值的平均值[21], 即

$$< k >= \frac{\sum_i^N k_i}{N} \quad (1)$$

其中, N 为传播网络中美国州的个数. 传播网络的平均度越大, 说明网络在传播过程中州与州的互动比较频繁.

3.2 度异质性

复杂网络的度异质性代表网络度的均匀程度, 对网络的性能具有重要的影响[22], 比如有文献指出, 在流行病的传播过程中, 度异质性越强, 流行病越容易爆发[23]. 在问答社区的传播图中, 度的异质性[24]可定义为:

$$H = \frac{< k^2 >}{< k >^2} \quad (2)$$

其中, $< k^2 >= \sum_{i=1}^N k_i^2 / N$, $< k >^2 = (\sum_{i=1}^N k_i / N)^2$.

由于问答社区的传播图是有向图, 传播网络的度值可分为入度和出度, 其中入度代表其它州指向该州的边的数目, 记为 k_{in} ; 出度代表该州指向其它州的边的数目, 记为 k_{out} . 我们把上式中的度分别改成入度和出度, 就可以得到入度异质性和出度异质性, 即

$$H_{in} = \frac{< k_{in}^2 >}{< k_{in} >^2} \quad (3)$$

$$H_{out} = \frac{< k_{out}^2 >}{< k_{out} >^2} \quad (4)$$

3.3 平均聚类系数

网络的聚类系数是用来描述网络节点聚集程度的系数 [25, 26], 如在社交网络中, 聚类系数可以描述一个人的朋友间相互之间的认识程度. 在问答网站传播图中, 与美国州 i 相连的州可用该州的度 k_i 表示, 这 k_i 个州之间实际存在的边数用 E_i 表示, k_i 个州之间最多可能存在的边数为 $k_i(k_i-1)/2$, 那么美国州 i 的聚类系数 C_i 可定义为

$$C_i = \frac{E_i}{n_i(n_i-1)/2} \quad (5)$$

平均聚类系数表示与同一个节点相连的两个节点也相连的平均概率. 传播图的平均聚类系数 $< C >$ 定义为

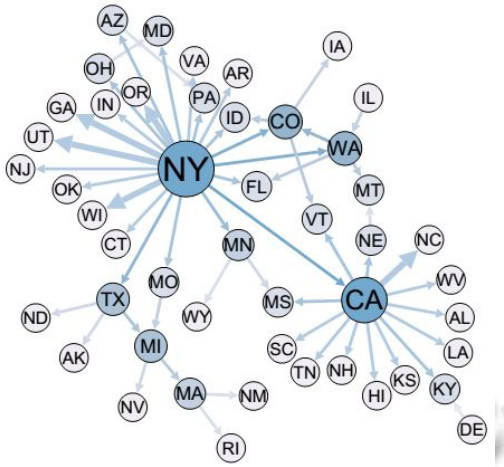
$$< C >= \frac{\sum_{i=1}^N C_i}{N} \quad (6)$$

3.4 网络直径

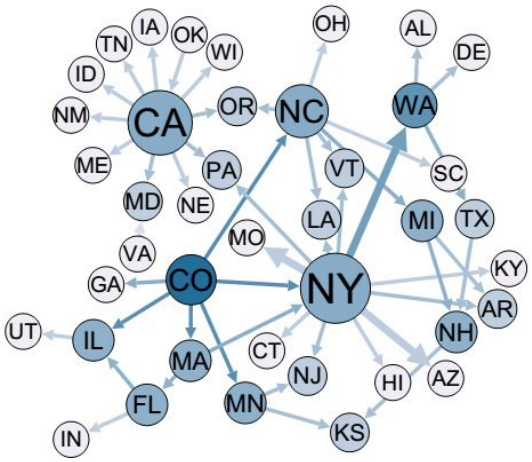
网络的直径表示任意两节点距离的最大值[26, 27]. 传播图中美国州 i 和美国州 j 的路径是指从美国州 i 开始, 经过与它相连的节点, 到达州 j 所经历的连边数. 问答网站传播图的网络直径 D 可以定义为:

$$D = \max_{i,j} d_{ij}, \tag{7}$$

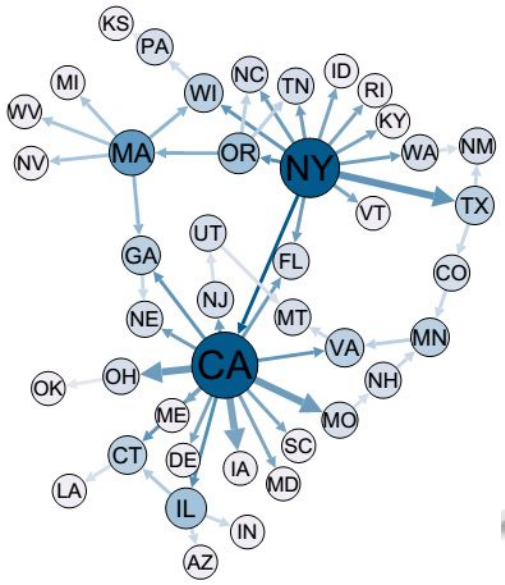
其中 d_{ij} 表示在美国州 i 和州 j 相连的情况下, 两州之间最短路径的边数[21].



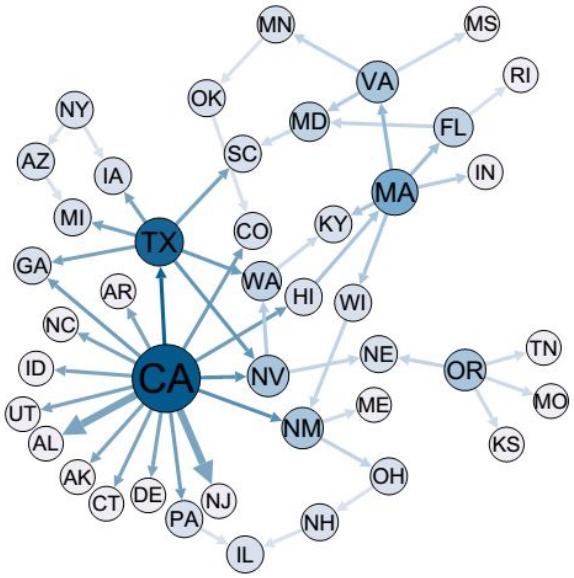
(a) Ask Different 社区传播网络



(b) Information Security 社区传播网络



(c) Cross Validated 社区传播网络



(d) Unix & Linux 社区传播网络

图 4 StackExchange 部分问答社区的知识传播网络

3.5 平均路径长度

网络的平均路径长度表示网络中节点间的平均分离程度. 传播图的平均路径长度 L 指的是任意两个州之间最短路径的平均值[21], 即

$$L = \frac{2}{N(N-1)} \sum_{i<j} d_{ij}. \tag{8}$$

4 结果与分析

4.1 问答社区的演化分析

知识网络的传播源应具有自我驱动能力强, 传播范围广的特点. 因此在传播图中, 我们把出度与入度的比值大于 5 的州作为传播源. 值得注意的是, 在此定义下可能出现入度为 0 的情况, 在此情况下, 我们定义传播源的出度与入度的差值不小于 5. 通过分析 23 个问答网站的社区传播图我们可以发现: 1)绝大多数传播图的传播源都不止一个, 这说明从问答网站的成立之初起, 随着时间的推移以及用户不断的参与问答活动, 问答网站的演化并不是单纯从一个州开始慢

慢向其它州进行传播, 而是从多个传播源开始向其它州进行传播, 如图 4(a)中的 Ask Different 社区, 其传播源是纽约州(NY)和加州(CA), 又如图 4(d)Unix & Linux 社区的传播源分别为加州(CA)、德克萨斯州(TX)和马萨诸塞州(MA). 相比于单个传播源, 问答网站的这种传播模式可能可以使知识更快地传播出去; 2)对比各图的节点大小和颜色深浅, 我们可以发现加州明显与其它州的互动更加频繁, 其它州的用户更倾向于与加州的用户交流沟通, 这可能与该州发达的经济和教育密不可分.

表 1 23 个问答社区的统计特征值

社区名称	$<k>$	H_{in}	H_{out}	$<C>$	D	L
Ask Different	1.167	1.1690	8.701	0.321	7	2.975
Ask Ubuntu	1.122	1.222	4.247	0.063	8	3.898
Code Review	1.273	1.207	4.506	0.127	8	3.909
Database Administrator	1.043	1.289	2.977	0.047	11	4.636
Drupal Answers	1.091	1.157	4.886	0.076	9	3.804
Electrical Engineering	1.114	1.120	4.565	0	8	3.896
English language & Usage	1.319	1.149	5.634	0.191	8	3.246
Arqade	1.320	1.192	3.543	0.005	7	3.376
Geographic Information Systems	1.224	1.252	4.749	0.154	7	3.682
Mathematics	1.229	1.2000	6.420	0.076	8	3.548
MathOverflow	1.325	1.152	7.483	0.224	6	2.858
Meta	1.460	1.149	5.393	0.226	6	3.096
Physics	1.093	1.226	4.575	0.051	8	3.853
Programmers	1.292	1.223	3.575	0.088	9	3.691
Science Fiction & Fantasy	1.405	1.170	3.820	0.194	6	3.150
Information Security	1.195	1.183	6.209	0.021	7	3.357
Server Fault	1.180	1.225	3.018	0	10	4.175
SharePoint	1.116	1.130	5.000	0.067	8	3.697
Cross Validated	1.238	1.149	6.720	0.210	6	3.059
Super User	1.265	1.249	3.128	0.071	8	3.718
Tex-LaTex	0.971	1.353	5.935	0	8	4.293
Unix & Linux	1.209	1.209	6.261	0.068	8	3.494
Wordpress	1.143	1.224	6.141	0.243	6	3.925

4.2 传播图的特征值统计与分析

表 1 给出了 23 个问答社区的 6 个统计特征: 平均度、入度异质性、出度异质性、平均聚类系数、网络直径、平均路径长度. 我们发现绝大多数问答社区的出度异质性远大于入度异质性, 这表明美国各州在传播过程中, 主要由若干州直接向其它州进行蔓延, 具有较强的直接传播特性. 不同问答社区间的平均度、平均聚类系数、平均路径长度以及网络直径差别不明显, 说明大部分传播网络结构相似.

5 结论

在线问答网站作为社交网络的一种, 在传播和分享知识方面发挥着巨大的作用. 研究这类网站的知识传播模式有助于我们进一步地了解在互联网上用户是如何分享知识的. 在本研究中我们研究了 StackExchange 中的 23 在线问答网站, 利用用户的问答关系, 构建出美国境内基于问答社区的知识传播图谱, 并探究不同问答社区的知识传播模式, 发现如下几个特点: ①多数问答网站具有多个传播源; ②经济和教育发达地区的用户交互更为频繁; ③不同知识领域的传播网络结构相似, 具有内在一致性.

参考文献

1 Alsina EF, Rand W, Lerman K. The success of question answering communities: How diversity influences ad hoc groups. Collective Intelligence, 2015: 1–4.

2 左美云,姜熙.中文知识问答分享平台激励机制比较分析-以百度知道、腾讯搜搜问问、新浪爱问知识人为例.中国信息界,2010,(11):25–30.

3 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).

4 Ahmed S, Yang S, Johri A. Does online Q&A activity vary based on topic: A comparison of technical and non-technical StackExchange forums. Proc. of the Second (2015) ACM Conference on Learning Scale, 2015: 393–398.

5 Correa D, Sureka A. Integrating issue tracking systems with community-based question and answering websites. Proc. of the 2013 22nd Australian Conference on Software Engineering. 2013. 88–96.

6 Bosu AS. Corley C, Heaton D, Chatterji DC. Carver JA. Kraft N. Building reputation in StackOverflow: An empirical investigation. Proc. of the 10th International Working Conference on Mining Software Repositories. 2013. 89–92.

7 Bazelli B, Hindle A, Stroulia E. On the personality traits of StackOverflow users. Proc. of the 29th IEEE International Conference on Software Maintenace. 2013. 460–463.

8 Movshovitz-Attias D, Movshovitz-Attias Y, Steenkiste P, Faloutsos C. Analysis of the reputation system and user contributions on a question answering website: StackOverflow. Proc. of 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. 2013. 886–893.

9 Morrison P, Murphy-Hill E. Is programming knowledge related to age: An exploration of StackOverflow. Proc. of the 10th IEEE Working Conference on Mining Software Repositories. 2013. 69–72.

10 Wang SW, Lo D, Jiang LX. An empirical study on developer interactions in StackOverflow. Proc. of the 28th Annual ACM Symposium on Applied Computing. 2013. 1019–1024.

11 Asaduzzaman M, Mashiyat AS, Roy CK, Schneider KA. Answering questions about unanswered questions of StackOverflow. Proc. of the 10th Working Conference on Mining Software Repositories. 2013. 97–100.

12 Treude C, Barzilay O, Storey M-A. How do programmers ask and answer questions on the Web. Proc. of the 33rd International Conference on Software Engineering. 2011. 804–807.

13 Xia X, Lo D, Wang XY, Zhou B. Tag recommendation in software information sites. Proc. of the 10th Working Conference on Mining Software Repositories. 2013. 287–296.

14 Vasilescu B, Capiluppi A, Serebrenik. A gender, representation and online participation: a quantitativ study of StackOverflow. Proc. of 2012 International Conference on Social Informatics. 2012. 332–338.

15 Posnett D, Warburg E, Devanbu P, Filkov V. Mining stack exchange: Expertise is evident from initial contributions. Proc. of 2012 International Conference on Social Informatics. 2012. 199–204.

16 MacLeod L. Reputation on StackExchange: Tag, you're it!. Proc. of the 28th International Conference on Advanced Information Networking and Applications Workshops. 2014. 670–674.

17 Furtado A, Andrade N, Oliveira N, Brasileiro F. Contributor profiles, their dynamics, and their importance in five Q&A Sites. Proc. of the 2013 Conference on Computer Supported Cooperative Work. 2013. 1237–1252.

18 Lal S, Correa D, Sureka A. MiQs: Characterization and prediction of migrated questions on StackExchange. 21st Asia-Pacific Software Engineering Conference (APSEC 2014). 2014.

19 Thongtanunam P, Kula RG, Cruz AEC, Yoshida N, Ichikawa K, Iida H. Mining history of gamification towards finding expertise in question and answering communities: Experience and practice with StackExchange. The Review of Socionetwork Strategies, 2013, 7(2): 115–130.

20 Schenk D, Lungu M. Geo-locating the knowledge transfer in StackOverflow. Proc. of the 2013 International Workshop on Social Software Engineering. 2013. 21–24.

21 Albert R, Barabsi AL. Statistical mechanics of complex networks. Reviews of Modern Physics, 2002, 74(1).

22 Hu HB, Wang XF. Unified index to quantifying heterogeneity of complex networks. Physica A: Statistical Mechanics and its Applications, 2008, 387(14): 3769–3780.

23 Wang W, Tang M, Zhang HF, Gao H, Do YH, Liu ZH. Epidemic spreading on complex networks with general degree and weight distribution. Physical Review E, 2014, 90(4): 042803.

24 Xuan Q, Du F, Wu TJ, Chen GR. Emergence of heterogeneous structures in chemical reaction-diffusion networks. Physical Review E, 2010, 82(4): 046116.

25 Watts DJ, Strogatz SH. Collective dynamics of 'small world' networks. Nature, 1998, 393(6684): 440.

26 汪小帆,李翔,陈光荣.复杂网络理论及其应用.北京:清华大学出版社,2006.

27 Boccaletti S, Latora V, Moreno Y, Chavez M, Hwang DU. Complex networks: Structure and dynamics. Physics Reports, 2006, 424(4-5): 175–308.