中国科学技术大学

硕士学位论文

社区问答系统中的社团发现技术研究及其应用

作者姓名：冯晓楠

学科专业：计算机软件与理论

导师姓名：田野 副教授

完成时间：二〇一四年四月

田野

2014-5-25

University of Science and Technology of China

A dissertation for master degree

Research of Community Detection in Community-based Question and Answering Systems

Author : Xiaonan Feng

Speciality : Computer Software and Theory

Supervisor : Associate Prof. Ye Tian

Finished Time : April, 2014

中国科学技术大学学位论文原创性声明

本人声明所呈交的学位论文，是本人在导师指导下进行研究工作所取得的成果。除已特别加以标注和致谢的地方外，论文中不包含任何他人已经发表或撰写过的研究成果。与我一同工作的同志对本研究所做的贡献均已在论文中作了明确的说明。

作者签名：汤晓丹　　签字日期：2014.5.25

中国科学技术大学学位论文授权使用声明

作为申请学位的条件之一，学位论文著作权拥有者授权中国科学技术大学拥有学位论文的部分使用权，即：学校有权按有关规定向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅，可以将学位论文编入《中国学位论文全文数据库》等有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。本人提交的电子文档的内容和纸质论文的内容相一致。

保密的学位论文在解密后也遵守此规定。

公开　　保密 _____ 年

作者签名：汤晓丹　　导师签名：

签字日期：2014.5.26　　签字日期：2014-5-25

摘 要

社区问答系统（Community-based Question and Answering System，CQA）通过聚集大众智慧，能够免费提供问题的个性化解决方案，例如 Yahoo! Answer，百度知道等。然而 CQA 系统无显式的社团结构，因此“社团”性质没能得到充分应用；而且 CQA 系统具有较高的开放性：知识内容共享和搜索引擎可接触，使得 CQA 系统易受到虚假账户的入侵，导致 CQA 账户行为规律复杂，知识质量急剧下降。

为解决 CQA 系统的上述问题，有必要深入研究系统中账户行为规律和网络性质。同时这些研究工作也有助于解决如下问题，例如相关用户推荐，相似问答内容融合，新型话题发现，虚假用户识别，个性化问答服务等，这些都能提高 CQA 系统中的知识质量。

本文以中国最大的 CQA 系统“百度知道”为代表，分析 CQA 系统中账户的行为规律。通过探索账户间的问答关系，本文构建两种网络模型，展示了 CQA 系统的基本网络性质。为检测 CQA 系统中的以兴趣为中心的账户社团，基于标签传播算法 SLPA，我们提出一个面向 CQA 系统的社团发现算法 MSLPA （Multilayer speaker-listener label propagation algorithm）。本文从网络规模、社团主题、聚合效果、层次结构等多方面评估 MSLPA 算法的性能，和已有的几种社团发现算法相比，MSLPA 能够发现大规模 CQA 网络中有意义的、重叠的、具有层次结构的账户社团，避免生成大量的微型社团，有效聚合关联账户。

基于 MSLPA 社团发现技术，本文提出一个 CQA 系统中鉴别虚假账户的方法。首先给出一组具有较高区分度的账户属性集合，包括具有一定物理含义的账户个体属性和账户所属的社团性质，其中个体属性由统计分析得到，社团性质由本文的社团发现结果得到。本文将新提出的属性集合应用于简洁的 J48 决策树分类器上，判断账户为正常账户或者虚假账户。实验结果显示，该方法表现出良好的性能和效果，分类准确率得到较大的提高。

关键词： CQA 系统，社团发现，MSLPA 算法，社会网络分析，虚假账户鉴别

ABSTRACT

Community-based Question and Answer (CQA) systems , such as Yahoo! Answer, Baidu knows, provide people with free answers to questions by integrating public intelligence. However, CQA systems have no explicit "community" structure, which would play an important role in many applications. Furthermore, CQA systems are vulnerable to spam accounts farming, because CQA systems can be accessed by Internet users and search engines, which decreases the quality of knowledge in CQA systems rapidly.

Considering the above problems, it is necessary to analyze characteristics of CQA users' behaviors and user networks. In the meanwhile, researching results can make contributes to handling the following problems, such as detecting spam accounts, providing personalized services etc.

This thesis focuses on "Baidu Knows", the largest CQA system in China, analyze CQA users' behaviors, and studies its social network structures. By exploiting the question-answer interactions among the users, two networks are constructed and shown to have strong social network characteristics. In addition, interest-oriented user communities can be observed on the networks. Furthermore, we propose Multilayer Speaker-listener Label Propagation Algorithm (MSLPA), an improved variation of SLPA, to detect user communities in CQA networks. MSLPA's performances are evaluated from aspects of network size, community topics, clustering, and hierarchy. Comparing with existing algorithms, MSLPA can effectively detect the genuine, overlapping, hierarchical communities in which users share common interests, and avoid forming large number of tiny communities.

Community detection technology is applied in identifying spam accounts in CQA systems. To detect spam accounts, we propose one group of account's

properties having high discrimination: including accounts' individual properties computed by statistical analysis, and accounts' community structure computed by community detection. By applying proposed properties in facile J48 classifier, experimental results show that these nice properties have good performance and the classification accuracy is improved.

Keywords: Community-based Question and Answering System(CQA),Community Detection, MSLPA, Social Network Analysis, Spammer Detection

目 录

表格

插图

算法

第一章 绪论

1.1 引言

近些年来，互联网发展十分迅速。Web2.0 技术导致了社区问答系统（Community-based Question and Answering Systems，CQA）的涌现和流行。CQA 作为一种新的知识分享和用户社交的平台，通过集合大众智慧，为人们免费提供问题的解决方案。社区问答系统日益发展壮大，知识数据越来越多，种类也日益繁多。

文献 [1] 定义问答系统为一种可以回答自然语言的自动机，从最初的专家系统发展到问答客服机器人系统 [2], 再到开放的社区问答系统以及在线社交媒体中嵌入的社交问答系统，能够为互联网用户提供针对性的问答服务，发展得更为迅速和智能。其中社区问答系统从 2005 年开始兴起，并快速发展成为知识搜索的重要组成，与以往的机器人问答系统相比，CQA 系统的知识库十分庞大，并且互联网用户也可以参与答案的评判。我们熟知的社区问答系统包括百度知道、雅虎知识堂、新浪爱问、腾讯问问等 [3]。社区问答系统根据话题覆盖面分类两种，一种是包含方方面面话题的广泛问答系统：Yahoo! Answers, 百度知道，Facebook Questions，另一种是仅针对单一或某几个少数主题的垂直型问答系统：stack overflow，知乎，Quora，搜房问答等。

社区问答系统可以弥补搜索引擎的不足，能够针对性地给出用户需求的解决方案。尽管搜索引擎持续发展了 20 年，仍然有许多用户的需求没能解决，主要原因是用户的潜在需求难以被完全表示出来，或者网络中缺乏对需求的解决方案。众所周知，很多用户的需求相对主观或者狭窄，甚至有些需求只能先有问题才能有解决答案，对于上述这些需求，不论搜索引擎发展到什么阶段，都难以给出较好的解决方案。相反，社区问答系统可以解决个性化的和开放性的问题，为用户需求创造与时俱进的解决方案。

国内的 CQA 系统具有类似的信息组织，表1.1 给出了 CQA 系统的组成主

体和主体的公开基本属性。国内 CQA 系统中的“百度知道”依赖于百度搜索，成长较快，有最大的用户群和搜索价值，所以本文以百度知道作为 CQA 系统的代表，简要介绍 CQA 系统的组织架构和知识流通过程。

“百度知道”系统将问题类别划分为“电脑/网络”、“生活”、“医疗健康”、“体育 \ 运动”等 14 个一级分类。然后一级类别进一步或者更多步细分为 917 个最小级别的分类，例如“电脑/网络”分为“硬件”、“常见软件”、“编程语言”等。用户提交问题时，可以自己选择问题类别，或者接受“百度知道”自动推荐的问题类别。一个问题可以被一个或多个回答者回答，提问者有权利选择其中的一个回答作为“最佳答案”。如果提问者需要问题较快得到回应或者希望得到质量较高的回答，可以通过增加悬赏分来吸引回答者关注。用户登录、提问或者回答等操作，都会带来不同程度的“收益”和“报酬”。用户回答问题会获得相应的经验值，如果问题被选为最佳答案，会获得财富值及对应的悬赏分。其中回答数和提问数反映了该用户对知识共享平台的贡献度，经验值反映了用户的活跃度，而财富值、赞同数、采纳率均反映了该用户的贡献的知识质量。问题类别定义了该问题的知识领域，悬赏分反映了提问者对问题答案的急切程度，浏览次数反映了该问题在互联网用户中的流行度，而问题的回答总数则在某种程度上反映了互联网用户掌握该问题知识的广泛度。

表 1.1 CQA 系统中主体的基本属性

类型	属性						
用户	ID	回答数	提问数	经验值	财富值	采纳率	赞同数
问题	ID	类别	提问者	提问时间	悬赏分	回答数	浏览数
回答	ID	回答者	问题 ID	回答时间	赞同数	是否最佳	是否匿名

CQA 系统能够通过多种途径促进用户问题的解决，例如从社区现有的知识库中提取信息，把问题广播给整个社区，将问题定向路由到朋友，或者为用户推荐相关话题的专家。然后实际的 CQA 系统一般只是列出待解决的问题列表，没有推荐机制。CQA 系统服务着数亿个用户和问题，持续产生大量的知识内容，已经积累了很庞大的知识库。本课题获取到的数据显示，截止到 2012 年 5 月，中文社区问答系统“百度知道”一共累计包含 332,834,000 多个问题，注册,6,240,00 多个用户，超过一半的问题解决时间在 30 分钟以内，这些数据表明

CQA 可以针对性地解决绝大部分用户的问题，及时性也较高，同时超过 25% 的搜索结果次数中 CQA 网站页面被列在搜索结果首页。

CQA 系统相对于互联网的其他社交媒体，例如 Facebook、人人网、微博等，在内容构成和功能上有着重大差异，因此有必要把 CQA 系统作为一种独立的研究对象进行深入研究。首先社区问答系统根据知识内容组织成树形的类别架构，例如百度知道中的一级问题分类“电脑 \ 网络”，拥有“电脑装机”、“互联网”、“程序设计”等九个二级分类，其中互联网又分为“上网帮助”、“网站推荐”等三个三级分类，而人人网和微博等则呈现一种平面式的内容结构，用户和用户之间自发形成关系链，不存在类别结构。其次社区问答系统不同用户的连接是通过问题的问答来建立的，用户难以发现具有相似兴趣的其他用户，也难以寻找到自身在整个用户群体中的定位，但是同时又会受到 CQA 系统中附加的其他社交功能的影响，例如私信功能、向他（她）提问等，从而使得 CQA 系统中的关系网络并非是单纯以话题为中心的信息网络；相比而言，人人网用户的连接则是直接通过“好友”关系建立。社区问答系统可以给用户解决十分个性化的问题，可以成为用户展现自身知识量的平台，是一种重要的在线学习方法，而人人网等则侧重于娱乐和社交。社区问答系统凝结大众智慧，对知识进行有效的普及和传播，扮演着重要的角色。

CQA 系统有较高的实用性和流行度，因此理解 CQA 系统中用户的行为规律，增加 CQA 系统的活跃度，检测 CQA 系统的虚假信息，保证 CQA 系统健康发展都有重要的研究价值。

1.2　CQA 系统的相关研究工作

本节给出国内外针对 CQA 系统的相关工作和研究不足。

到目前为止，国内外针对社区问答系统有相当丰富的研究工作和成果 [4–9][10]，研究方向包括专家检测 [5, 11]，用户行为模式分析 [8, 12]，问题质量测量 [13–17]，答案质量评价，问题推荐 [18, 19]，问题路由，虚假用户检测 [20, 21] 等。

社区问答系统涉及到较多的自然语言处理，因此针对问答内容和质量的研究工作比较丰富。文献 [10, 15] 利用关键词和问题统计属性判断问题类别，文

献 [4, 7] 利用已有的可靠样本和回归分析，定性和定量的衡量问题质量。文献 [14] 为了提高 CQA 系统中新问题的解决速度，基于问题自然语言的相似性，从系统中过去的问答中提取答案推荐给新问题，提出了一套基于社区问答系统知识库的在线机器人问答算法。文献 [18] 为了让问题得到尽快解决，应用问题和用户的局部信息，以及问题类别集合中的全局信息，使用 SVM 分类的机器把问题路由到对应的用户。文献 [16] 将 CQA 中的问答内容与维基百科的概念进行相似度分析，为 CQA 系统中新涌现的话题自动贴话题标签。

1.2.1 CQA 系统中的账户行为研究

国内外研究机构不仅深入研究了 CQA 系统中的问答内容和语义方面，而且在对用户的行为规律上也有一些研究成果。文献 [9] 利用 Yahoo! Answers 一个月的问答数据，构建从提问者指向回答者的有向图，分析不同问题类别间的网络属性差异，对用户参与的多个问题类别组合进行关联分析。文章发现不同类别的问题，在内容和网络结构上均有明显的差异，而用户经常活跃的问题类别之间也有关联性。文献 [22] 对 Naver Knowledge iN 问答社区详细分析了用户在不同领域的表现，问题提问者设置最佳答案的倾向，问题参与者的活跃时间间隔等。文章旨在理解用户的行为规律，从而改进社区问答系统机制来适应用户习惯。文献 [23] 针对 stack overflow 网站，基于问答和用户的属性，利用逻辑回归算法预测一个问答页面的参考价值，发现问题提出后的前 72 个小时发生的交互时间能够精确预测该问答的知识价值。文献 [8] 同样利用 stack overflow 数据集，建立正常用户和专家用户属性变化的时间序列，使用聚类算法发现专家用户属性的演化模式，并将发现结果运用到专家用户的鉴别中。可以看出国外 CQA 系统更加流行，因此研究成果颇丰，然而对国内社区问答系统的结构研究仍然比较欠缺。文献 [24] 使用统计方法研究新浪微博中影响社交问答 (social Q&A) 回复率的因素，包括提问者粉丝数，提问频率，问题分类等。文献 [25] 收集 Yahoo! Answer 中新用户最初一星期的问答数据，使用基本分类算法鉴别 CQA 系统中新用户是否有打算停止使用 Yahoo! Answer 的倾向。文献 [5, 13, 26] 讨论问答系统中不同参与者群体的功能和行为偏向，文献 [6, 27] 研究问答系统背后和人文相关的因素，理解人们参与回答的动力。文献 [10] 研究了百度知道的问题类别鉴别，但是没有考察用户的行为特征，即国内还没有在百

度知道用户行为方面的研究工作。

调研发现大部分研究工作在考察用户的行为规律时仅仅使用到用户的个体信息，忽略了用户在系统中的全局特征，然而 CQA 账户的个体属性不足以表现出 CQA 系统中账户的交互行为规律。CQA 账户通过问答行为形成连接关系而构建的网络模型称为“CQA 账户网络”，CQA 账户网络能体现 CQA 系统账户的交互行为和全局状态。相对于互联网的其他社交媒体，例如 Facebook、人人网、微博等，CQA 系统没有明显的“好友”关系，因此如何构建合理的 CQA 账户网络成为一个重要的课题。文献 [12] 中构建的有向网络可以反映出专业信息和知识的供求方向，例如一个从提问者指向回答者的有向网络，其中“专家”用户的入度要远远大于其他用户的入度，但是有向网络不能反映出回答者之间会相互参考点评交互的行为，不易于直接表现相关用户的交互和聚集。据我们所知，只有很少的工作关注 CQA 账户网络，因此本文在对用户个体分析的基础上，进一步使用社交网络分析方法研究用户的行为。

在现实生活中，人们有直观的社交行为，可以和朋友形成一个“社团”，或者几个具有不同含义的“社团”，例如“同学社团”，以美食为话题的社团等。一个社团是指由个体组成的一个功能性系统，社团内个体之间的交互要比个体与群体外的交互更加频繁。相比较，社交媒体为人们在网上提供了像在现实生活中社交一样的平台，而且不受地理距离和成长经历的限制。在线社交网络中社团的形成主要有两种方式，一种是显式社团，依赖用户自身订阅朋友，通过申请添加好友形成，例如微博中的同事群组，好友圈等；另一种是隐式社团，通过用户个体间的交互行为反映出来，例如 CQA 账户网络。社团发现技术可以应用到很多方面，例如简化视图的可视化，分析用户节点行为规律，在搜索引擎中用于分类，追溯社团演化规律，探索功能性单元组成，分离大型社区等。在社交媒体形成的大规模用户网络中检测用户社团成为了当今研究的热点之一 [28–31]。

到目前为止，有较多的研究团队致力于社交网络中的社团发现技术 [30, 31]，例如 CPM(clique percolation method)[32], 随机游走 [33]，谱聚类 [34]，模度最大化 [35, 36] 等算法。文献 [28] 通过在用户网络中添加节点信息和边信息，应用 k-means 聚类算法发现社团, 并将发现到的社区和用户订阅的好友群组进行对比。文献 [29] 则给出了能够处理百万个视频的社团发现框架，基于用户观看

的视频记录和视频相似度，并行计算密度划分算法获取社团，然后从社团成员的交互行为中抽取社团主题。

根据网络中发现的社团之间是否具有共同成员，可以将社团种类分为分离性社团和覆盖性社团，其中前者社团之间没有共同成员，后者是指社团之间存在成员交集，即某个成员可能存在于多个社团中。较多文献给出了不同思路的重叠型社团发现算法 [32, 37–39]，OSLOM[39] 算法是基于局部信息的最优化算法，通过计算在随机网络中出现某种特定社团的概率，并设定阈值来选择较为显著的社团；SLPA[38] 算法分辨率相对较大，复杂度和网络的边成线性增长，更适合百度知道较松散的大规模网络结构，然而下文的实验发现 OSLOM 虽然具有发现层次社团的功能，但是不适合于稀疏网络；而 SLPA 却没有检测社团层次性的功能，且对同类账户间的聚集力度不够。本节为重点分析 CQA 账户的社团聚集性质，提出一种针对 CQA 账户网络特性的社团发现算法。

以问答关系形成的 CQA 账户网络蕴含着大量的科研和实用价值，例如检测发现网络中的隐式社团，话题挖掘，问题推荐，用户推荐等。社团发现技术有助于我们更好的理解复杂网络，分析网络结构，甚至优化大数据存储。与其他社交网站（例如文献 [30] 中的 Flickr）不同，CQA 系统和 Youtube 相似，没有用户订阅的显式社区，这使得挖掘其中的隐式社区更具有重要意义，基于社区发现结果，我们可以进一步推荐相关用户解答问题，或者发现新型话题等。社团分析可以用来为百度用户推荐具有相似兴趣或者相同问题的其他用户，使得百度知道用户可以针对性的问答，提高用户满意度，促进知识的有效传播。

1.2.2 CQA 系统中虚假账号的植入行为研究

本文在调研实际的 CQA 系统时，发现一个日益明显的“虚假信息注入”的问题 [40]。在互联网上，部分网民探索出一种新的网上工作方式，他们在不同的在线网络社区和网站上张贴评论和文章，并得到费用，本文称这些网民为付费张贴者。然而从付费张贴者获得的信息通常是不可信的。由于 CQA 系统拥有大量的用户，且知识库可以在搜索引擎的结果中展示并获取流量，因此 CQA 系统成为网络营销的重要战地。CQA 系统作为一种知识分享和用户社交的平台，这些垃圾信息严重影响了问答系统的生态环境，也间接降低了大众的搜索质量。以“百度知道”为例，根据百度知道协议和“百度知道”的检举原

则，将垃圾问答分为四类：一是无意义灌水，例如“顶”、“同上”等无意义的回答，恶搞、人身攻击等性质的提问或回答；二是广告营销类，例如为增加流量而故意引导他们到某个网站或论坛，征友类问题，涉及求租转让的提问和回答；第三是色情类问答；第四是危害国家安全类问答。其中第一类和第三类可以通过自然语言处理技术得到良好的自动化检测。例如文献 [17] 基于用户和话题之间的交互关系与问答质量的关系，提出一种标签传播算法，根据交互数据预测问答质量；文献 [19] 使用机器学习的方法，通过组合优秀的学习算法为答案排序。

和第一类的低质量问答相比，第二类的营销问答目标是向受众推送特定的信息，例如某商品物美价廉，某网店货品丰富等，其中的问答的语义质量较高，问答的形式十分正式。这些伪装特性可以让营销问答很好的植入到 CQA 系统中，且总能吸引部分用户相信营销的内容。事实上营销问答可能会产生十分严重的后果，例如推销虚假药品，声誉并不好的教育机构，甚至还有钓鱼类网站的推广。这些营销问答批量地涌入 CQA 系统，使得参考价值和问答的真实性岌岌可危。

营销问答背后有着比较完整的资金链支撑。由于营销问答需要一定的规模才能在 CQA 系统中凸显出来，吸引用户，所以虚假信息发布者会在营销信息交易网站中发布任务。虚假信息交易模式如图1.1 所示：任务给出了问答内容和对应的报酬，很多个虚假信息注入者领取任务，在 CQA 系统中执行任务，张贴满足要求的问答。虚假信息发布者在审核注入者的任务后，给予对应报酬。虚假信息交易经常在猪八戒网、天才城等威客类网站进行。目前国内的各大 CQA 系统采用人工审核和人工删帖的方式来解决这种方法，例如“百度知道”相关规定中说明“百度知道”中垃圾问答主要依靠用户检举和管理员审查发现。考虑到这种方式成本昂贵，效率低下，有必要寻求一种自动化的鉴别方案，清除垃圾问答和恶意用户，改善问答系统的环境，提高网络用户的满意度。

本文将 CQA 系统中的账户根据发帖意图分为两类：正常账户和虚假账户。其中正常账户包括确实有需求而发布问题的提问者，和确实愿意提供与问题相符的答案的回答者；虚假账户指用户在发帖中另有所图，意欲欺诈的提问者和回答者。实际上威客在张贴“问答”内容时，清楚地知道自己并没有与“问答”内容相类似的经历，属于使用欺诈的方式来获取利益的不道德行为，而此类威

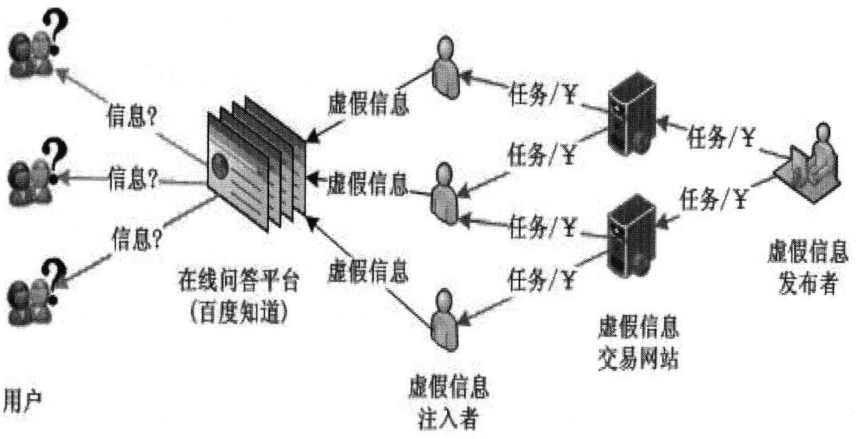

图 1.1 虚假信息交易模式

客属于虚假账户。实际中虚假账户数量居多，虚假信息注入量庞大，研究虚假信息的生存机理，并鉴别和阻止虚假信息的扰乱显得十分重要。

CQA 系统中的虚假账号和僵尸账号存在着较大差异。例如微博中的僵尸粉，这类账户可以通过程序批量产生，通过关注其他账号来增加其他账号的粉丝数，从而进一步增加关注账号的影响力，不需要培养僵尸账号的级别，几乎不需要交互行为。然而 CQA 系统用户之间没有“关注”或者“好友”等持久的关系，需要虚假账号不断地主动交互，例如发帖，回帖，才能生存下去达到营销目的。正是由于这种交互的伪装性，使得 CQA 系统中虚假账号的鉴别难度极大。文献 [41] 基于人人网账号间的好友申请和接受关系，构建社交网络，给出虚假账号的鉴别规则，发现虚假账号社团中绝大多数连接随机性的连入正常用户，而小部分连接其他虚假账户，并指出目前还没有有效的技术来组织虚假袭击。文献 [21] 研究了 twitter 的虚假账号生态系统，指出了虚假账号之间的交互关系，以及虚假账号支持群体，并设计出一种类 pagerank 的算法检测更多虚假账号。

1.3　本文的研究内容与主要贡献

经过对国内外研究现状的分析和实际 CQA 系统使用的调研，本文发现面向 CQA 系统的研究工作主要存在以下两个不足：

1. 由于 CQA 系统数据庞大，用户间无明显的好友关系，用户问答操作关系复杂，如何建立恰当的 CQA 系统的账户网络模型？另外基于 CQA 系统的网络模型，如何分析账户的网络属性特征和群聚性质？有哪种社团发现算法能够高效检测出有意义的社团？

2. 大批量虚假账号涌入开放性的 CQA 问答系统，肆意发布虚假信息，他们的欺诈行为严重影响了 CQA 系统的用户体验，然而 CQA 系统用户数据巨大，如何高效鉴别数亿用户中的虚假账号？

针对上述不足，本文利用网络爬虫程序爬取 CQA 系统中的真实数据，基于已爬取的数据分析 CQA 系统的用户和问题特征，构建账户网络模型，使用社会网络分析方法来描述 CQA 系统用户的行为规律，提出 MSLPA 社团发现算法进行社团聚类，并对发现的社团分析其组成成分同时将结果应用与虚假用户检测。本文提出的社团发现算法适用于大多数具有基本 CQA 系统主体的 CQA 系统平台，能够高效鉴别出合理的用户社团；虚假账户鉴别方法同样适应于大多数 CQA 系统平台，具有一定的普适性。

本文的主要贡献如下：

1. 设计一个面向 CQA 系统的网络爬虫系统，该系统在本文中应用于“百度知道”数据的抽取，然后通过修改单个文件，该爬虫系统不局限于“百度知道”，可以作为中文 CQA 系统数据爬去的通用爬虫。

2. 依赖获取的数据，分析 CQA 系统中账户行为的个体属性特征，同时提出新的账户属性，包括“回报率”，“类别熵”，“问答时间差”等，这些属性在鉴别虚假账户时具有较高的区分度。

3. 本文给出两种具有不同应用场景的 CQA 网络模型：问答完全网络和回答环形网络，并分析 CQA 网络的属性特征；进一步，本文提出了一个新颖的社团发现算法 MSLPA，该算法能够有效发现有意义的、重叠的、具有层次结构的账户社团。通过与已有的社团发现算法比较，MSLPA 适用于大规模网络，发现的社团质量较好。在 CQA 系统的庞大用户群中使用 MSLPA 算法发现用户社团，可以推荐相关用户，融合相似问答内容，发现新型话题，检测虚假用户等。上述应用均有助于用户获得针对性的解决方案，提高用户满意度，促进知识的有效传播。在面向中文 CQA 系统的账户行为研究中，本文首次给出并分析了具有不同应用场景的两种 CQA 账户网络模型，并提出一种面向 CQA 系

统、简洁高效、可扩展性强的社团发现算法。

4. CQA 系统的开放性使得其本身很容易受到虚假信息的植入和污染，而且虚假信息注入量十分庞大，作为社团发现的一个有效应用，基于具有区分度的 CQA 账户个体属性集合和社团发现结果，给出一个虚假账户鉴别的决策树分类方法，该方法简单高效，比不使用社团发现结果的分类准确率提高很多。本文给出的基于社团发现的虚假账号鉴别技术，能够阻止虚假信息的扰乱，具有一定的应用价值。

CQA 系统每天都在快速地增加新用户和问题，用户在 CQA 系统中贡献并享受所有的知识库，分析这些数据的内在结构和组织形式对于社会网络的研究和优化都有重要的意义。CQA 系统能够针对性地高质量地解决用户的问题，同时也给“专家”用户提供一个共享知识和给予帮助的平台，了解 CQA 系统中用户的行为规律，研究 CQA 系统中用户的行为规律显得十分有价值。能够更好更及时地提供 CQA 服务，使得发布需求的用户较快地获得高质量的回答，维护 CQA 系统的知识共享环境，抵制水军、广告、垃圾问答的侵入，促进 CQA 系统的健康发展。

1.4 论文的组织结构

本文主要分为以下几个部分：

第一部分讲述社区问答系统的发展背景和国内外研究现状，以及本文的研究内容概况和主要贡献。调研发现已有的研究工作中缺乏对中文 CQA 系统中的账户行为规律分析和网络属性研究，对 CQA 账户网络中的虚假账户研究更是少之又少。同时本部分简要给出了本文的研究概况：探索 CQA 系统中的用户行为规律，维护社区问答系统在开放的同时能够抵抗虚假信息的注入。

第二部分给出了面向社区问答系统的高效率爬虫技术，详细介绍爬虫系统的设计架构，爬取策略和执行方案，并给出爬取的数据概况，说明测量合理性。此外，为了后续分析，本部分抓取了著名网络营销信息交易网站“猪八戒”网站中有关“百度知道”的虚假信息，用以获取虚假样本。

第三部分揭示 CQA 系统中用户和问题两大主体的属性特征，对于每一个主体，本部分又分别从正常账户和虚假账户两方面来展示行为特征之间的差异，

包括两类用户的提问类别活跃度，问题解决率，答案数和浏览数的分布等。并进一步对行为差异给出合理的解释。在账户属性分析过程中，本部分还提出具有明显物理含义的账户属性，这些属性在下文的鉴别虚假账户时具有较高的区分度。

第四部分使用社会网络分析方法探索 CQA 系统账户的网络性质，包括网络模型，基本网络属性和社团特征。分析网络属性发现，CQA 系统中存在以用户兴趣为主题的账户社团。本部分提出了面向 CQA 系统的 MSLPA 社团发现算法，该算法尤其适用于大规模的 CQA 账户网络。将社团主题与账户自标记的主题进行对比，本章结果表明 MSLPA 算法能够有效发现有意义的合理社团。将该算法与已有的四种社团发现算法进行比较，实验结果表明 MSLPA 算法最适应大规模网络，且能发现层次性社团结构。通过分析 MSLPA 给出的不同网络层次结果，本章发现 MSLPA 能够有效融合社团，有效避免产生大量微小社团。

第五部分给出社区问答系统中虚假账户的鉴别方法。作为社团发现技术的一个应用，将社团发现结果应用到虚假账户鉴别中，给出了用于虚假账户鉴别的属性集合和最佳分类器类型。实验结果显示，该解决方案可以有效地在大规模 CQA 数据中鉴别出虚假用户。本章进一步分析账户类型误判的原因，这有助于挖掘出更多的虚假账户特征。

第六部分对现有工作进行总结，指出工作中存在的问题，并给出下一步的工作内容和进展方向。

此页不缺内容

第二章 CQA 系统的大规模数据爬取

本章给出了针对 CQA 系统的大规模网络爬虫技术，以“百度知道”为数据源，使用网络爬虫收集“百度知道”真实数据，同时，我们还从虚假信息交易网站中收集部分数据，用于分析虚假账号的行为模式。

2.1 CQA 系统的网络爬虫系统

2.1.1 CQA 系统的网络爬虫系统设计

网络爬虫是一个自动提取网页的程序，传统爬虫从一个或若干初始网页的 URL 开始，获得初始网页上的 URL，在抓取网页的过程中，不断从当前页面上抽取新的 URL 放入队列，直到满足一定停止条件。本节讲述网络爬虫的设计结构和执行策略，包括种子选择、HTTP 代理爬取、爬行策略和停止条件。表1.1给出了 CQA 系统中三大主体的通用的属性，这些属性将用于分析 CQA 系统的账户特征。

在“百度知道”中，每个账户会有一个账户主页，该账户主页包含表1.1中“用户”一栏的所有信息，以及用户的回答问题列表；问题页面的 URL 地址中，例如http://zhidao.baidu.com/question/461391986 ，数字串“461391986”称为“问题 ID”，问题 ID 号与问题页面 URL 一一对应。在问题页面中，包含“问题”和“回答”两栏信息，可以从问题页面进入到提问者或者回答者的账户主页。针对 CQA 系统的网络爬虫根据账户和问题之间的链接关系获取账号和问答数据。由于百度知道通过树形结构组合各类问题，为了涉及更多分类的问题和用户，本文通过基于深度优先的网络爬虫程序抓取百度知道数据。

考虑到“百度知道”问答系统拥有数亿的问题，爬取全部的问题是不现实的，因此需要采取抽样策略。常用的网络抽样方法有三种：节点抽样、连接抽样和雪球式抽样 [3]。节点抽样是指随机选取一定比例的节点，而节点之间的连接则全部选取。连接抽样是指随机选取一定比例的连接，然后和连接相关的节

点全部包含在最后的抽样网络中。雪球式抽样方法随机选取一些种子节点，然后进行宽度优先遍历，直到遍历到的节点数达到一定的抽样比率。然而三种抽样方法均严重依赖于初始种子的选择。本文采取与文献 [42] 中类似的遍历方法，在全局 CQA 系统中爬取数据，在爬取中随机选择下一步爬取方向，将文献 [42] 中的完整爬取和随机选取子集两个步骤统一在爬虫程序中。大多数已有的相关工作中数据集是网站的公开数据集，或者使用 API 获取几千个账户的完整数据，整体信息缺乏，然而每个账户问答次数较多，存在信息冗余。相比而言，本文仅爬取每个用户问答信息的部分子集，同时保证了整体数据规模较大，且范围较广泛。

图2.1详细介绍网络爬虫算法流程：为尽可能保证网络爬虫的覆盖性，本文从 14 个大类共收集 5,000 个问题 ID 作为种子，然后进入这些问题页面，分别解析问题和回答的属性，存入数据库中，同时获取到提问者和回答者形成账户列表。从账户列表中随机挑选一个用户，如果用户不曾被遍历到，则将用户的个人信息存入数据库。接着进入该用户的个人主页获取其回答问题列表，从问题列表中选取未曾解析过的问题，进行下一轮的问题解析，如此循环往复。根据百度知道信息组织结构和实验发现，深度优先遍历时重复进入同一问题页面的比率较低，有效爬取速度较快。

爬虫程序在全局的 CQA 系统中随机抽取问答信息，一个用户回答的问题越多，进入该用户主页的次数就越多。程序执行时间越长，获取的信息覆盖率越大。本文使用用户增加率来决定程序的停止时间，即每天在数据库中新增加的用户数。考虑到“百度知道”每天都会有新用户加入，实验设定当用户增加率在一定时间内低于某个阈值时结束爬虫程序，此时数据库中每个用户的问答记录与真实问答记录相比会同比例减少。因此最终获取的数据集中的问答信息是已爬取用户的所有问答信息的子集，但是对于一个问题收集了所有的回答信息。

2.1.2　网络爬虫系统执行策略

为保证针对 CQA 系统的爬虫程序正常运行，我们在访问方式和运行时间上都有所控制。考虑到“百度知道”反爬虫机制可能带来的影响和干扰，爬虫线程会轮流使用分布在全国各地的 HTTP 代理向“百度知道”发出访问请求。

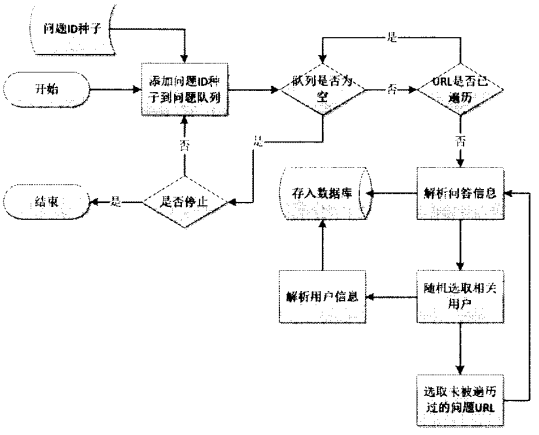

图 2.1 爬虫程序流程图

HTTP 代理列表会每日进行更新，同时检验代理工作是否正常，保证数据正常爬取。实验中使用到的 HTTP 代理为 100 个左右。爬虫程序一共运行 26 天，从 2012-05-03 至 2012-5-25 和 2012-5-31 至 2012-06-02，为减少对“百度知道”密集型访问给本地网络产生的影响，爬虫在每天的非高峰访问时间段运行，设定为 0：00 到 8：00。爬虫执行初始阶段，当同时开启 10 个线程的时候，每个小时增加 30,000 个用户，10,000 个问题，50,000 个回答。图2.2中横轴指爬虫系统运行的总时长，以小时为单位，纵轴指爬虫系统在对应的时间段中新爬取的用户数。由图2.2 可以看出新增用户数逐渐减少，当新增用户数减少为 3,000 左右时，我们停止了爬虫程序。

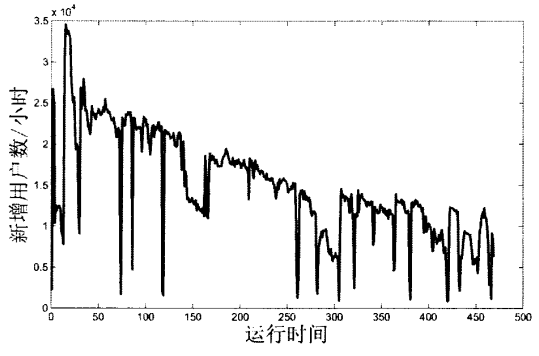

图 2.2 新增用户速率

爬虫一共爬取 6,979,013 个用户，4,992,514 个问题和对应的 24,600,297 个回答, 即收集到实际百度知道问题数目的 1.3%，实际百度知道回答数目的 1.5%。对于每个问题，爬虫会完整的记录问题和所有回答。初步分析数据，覆盖的问题类别包含了所有的一级问题类别和 99.1% 的最底层小类别，爬取的类别分布与百度知道整体的数问题类别分布相同，所以爬取结果比较均匀。在 2013 年 10 月下旬“百度知道”改版之前，问题 ID 数值有随着提问时间连续性增加的规律，分析“百度知道”问题的 URL 和提问时间，如图2.3所示，提问时间越靠后，问题 ID 的数值越大。已知数据库中最大的问题 ID 号为 432014465，已收集到的问题数目 4,992,514 与可能的最多问题数 432014465 相比约为 1.2%，与上文中的 1.3% 接近，稍微偏小可能是由于部分 ID 号对应的问题页面已被删除。

由图2.3可知问题 ID 数值有随着提问时间连续性增加，图2.4给出了数据库中问题 ID 数值的分布，和预期的均匀分布不同，其中在 1,000,000,000 左右出现较大的峰值。图2.3中可以看出在 2009 年左右曲线迅速上升，导数偏大，“百度知道”涌现出大量的问题。这可能是由于“百度知道”在 2009 年开始大范围流行起来，从而用户之间的问答关系变得十分稠密，因此本章提出的基于问答关系的爬虫程序，也相对密集地收集了该时间段的数据。

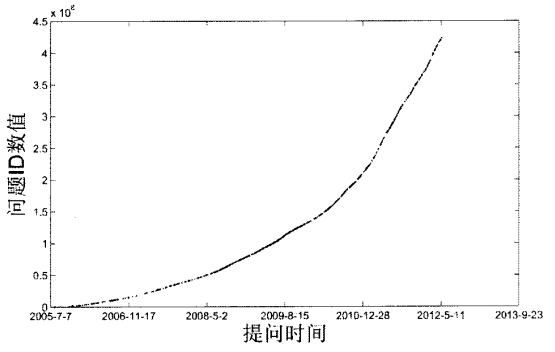

图 2.3 问题 ID 与提问时间的关系

分析爬取到的 8,000 个用户的相关数据，如图2.5所示，横坐标是指爬取的提问数占用户实际提问数的比例，纵坐标表示具有该种爬取比例的用户比例，可以看出，超过 70% 的用户的问题爬取比例集中在 3% 以下，回答爬取比例集

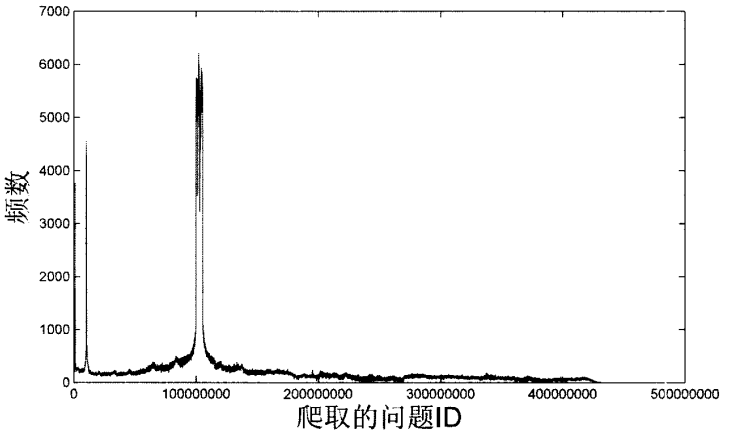

图 2.4 问题 ID 的数值统计

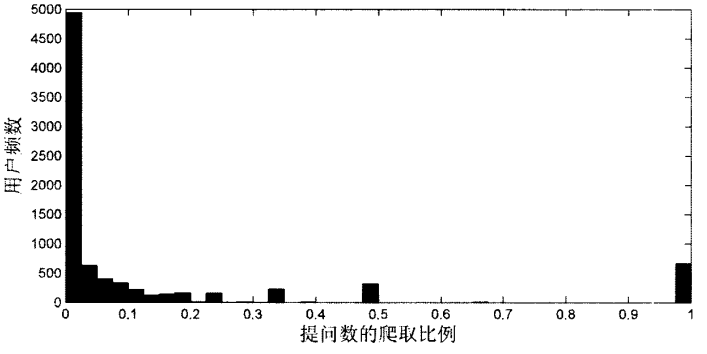

图 2.5 用户问题爬取比率

中在 10% 以下。

通过分析爬取的数据结果可以发现，本文设计的网络爬虫系统能够均匀抽取 CQA 系统的部分数据，并尽可能保证保留原始数据的特性。

2.2　CQA 系统中虚假账户的收集

第一章中将 CQA 系统中的账户根据发帖意图分为两类：正常账户和虚假账户。其中正常账户包括确实有需求而发布问题的提问者，和确实愿意提供与问题相符的答案的回答者；虚假账户指用户在发帖中另有所图，意欲欺诈的提问者和回答者。本节重点介绍一种高效准确的虚假账户的判别方法。

2.2.1 营销问答交易平台

威客网是一类典型的营销问答交易平台，一些拥有经验能力和时间的人群，在互联网上解决科学、技术、生活、工作中的问题，从而获取实际收益，例如威客任务中国，创意网，猪八戒网等。威客模式网站的用户可以分为任务发布者和回答威客。常见的交易模式为：任务发布者在威客类网站上发布任务，众多威客领取并参与任务，任务发布者审核后支付报酬给威客。

猪八戒网作为全国最大的在线服务交易平台，服务交易品类涵盖创意设计、网络营销、文案策划等多种行业。其中网络营销包含论坛推广、QQ 群推广、问答平台推广、推广之策、百度贴吧推广、百度知道推广等。例如任务发布者为了推销“石子进销存系统”这款软件，任务发布者会在猪八戒网发布如网址http://task.zhubajie.com/1270222/ 所示的任务详情，并给出问答示例：

> 1、发帖形式为一问一答（或多答），选中自己的答案为最佳答案，算一个有效稿件，每稿 1 元。
>
> 2、自问自答必须更换 IP 和账号，问题回答后必须超过三小时以上再选最佳答案，（最好是“今天发问题，回答，明天选最佳答案”，可防止问题被删除），问题被网站删除无效。
>
> 示例：问：想找个简单点的进销存软件用于服装店记账，谁用过类似的软件？答：《石子进销存系统》软件可以的，挺简单实用的，卖出的每种商品的成本、利润以及货品库存一清二楚，我的鞋店一直在用它管账，应该会适合你的

威客会根据任务要求在“百度知道”上按照示例进行问答，例如http://zhidao.baidu.com/question/347336345.html 是威客在“百度知道”上的一个问答页面。威客在进行问答后需要向任务发布者“交稿”，在猪八戒的任务页面张贴上述问答链接或者截图，以表示完成任务，然后任务发布者审查稿件是否合格，并进行计件，支付相应报酬。在这个过程中，威客并没有使用过“石子进销存系统”这款软件，属于使用欺诈的方式来获取利益的不道德行为。

为了获取“百度知道”中的虚假账户样本，从数据库中人工辨别虚假账号显得十分不现实，一方面账户数目比较庞大，另一方面没有明显的参考标准能够说明一个账号是否为虚假账户。鉴于此，本文采用从猪八戒网上交稿时贴出

的问答链接中抽取账号列表，这些账号规模较少，但是更有可能是虚假账户。同时我们对抽取出的账号列表进一步进行人工鉴别。需要注意的是，根据这种样本选取方法获得的虚假样本可能会存在一定的偏向性，例如会漏掉具有其他模式的虚假账号。但是这种缺失可以通过本文后续的社团发现方法进行修正。

从猪八戒网上爬取数据方法如下: 首先进入任务列表页面: http://task.zhubajie.com，选择包含“百度知道”关键词的任务，进入任务主页，获取任务 ID、任务内容和稿件。爬取程序从 2013 年 4 月开始运行执行 10 天，爬取从 2011 年 1 月到 2012 年 3 月的“问答网站推广任务，共收集到 2,293 个相关任务，一般一个任务都会有 20 个以上的稿件，而在“百度知道”上张贴的稿件中，合格稿件即合格的问题页面 9,258 个，不合格稿件的问题页面 3,094 个。通过人工检查和 2,293 个任务相关的问答页面，可以从页面中抽取账号列表，然后人工审查账号是否为虚假账号。

2.2.2 虚假账户标记

尽管上文定义虚假账户指用户在发帖中另有所图，意欲欺诈的提问者和回答者，然而不能依赖“虚假账户”的定义进行实际的鉴别操作，因为账户张贴内容的意图并不会被记录。本文中的“虚假账户”特指在 CQA 系统中发布广告营销信息的专用账户，将从猪八戒网上获取的相关 CQA 系统账户作为虚假账户的样本。

为了获取虚假账号的样本，本文从猪八戒网站爬取了 2,000 多个在“百度知道”张贴虚假内容的任务和对应 10,000 多个稿件，其中一个稿件为一个问题页面，包含提问者和回答者账号，我们希望从这些账号中鉴别出虚假账号作为先验知识。根据任务对威客的发帖要求，本文把任务分成如下三类模式，并给出每类模式中虚假账号的提取规则：

- 一问一答模式，即任务发布者要求威客用“百度知道”账号提问，过段时间再用另一个“百度知道”账号回答，再过段时间需要把该回答设为最佳答案。该种模式会对提问回答的内容给出完整的参考案例，方便威客执行此类任务。数据显示大约 94% 的任务属于该种模式。一问一答模式中我们假设问题的提问者和回答者均为虚假账号。我们在 2013 年 8 月份对此

类稿件再次爬取这些稿件，分析得知该种模式下稿件的存活率为 72%，即仍有 72% 的稿件没被“百度知道”管理员删除。

- “客服”模式：威客在问答系统中搜索包含营销对象的问题，然后按任务内容回复问题。该种模式中，我们将合格稿件中的回答者标记为虚假账号。大约 5% 的任务属于该类型，稿件存活率为 80%。
- “枪手”模式：威客只需要在问答系统中提出相关问题或需求，回答由任务发布者负责。该种模式中，我们将合格稿件中的提问者标记为虚假账号。这种模式的任务占 1%，由于问答操作复杂，更逼近于正常的问答模式，所以存活率达到 92%。

我们将任务人工分类标记，通过上述三种规则获得 9,218 个虚假账号样本。为避免判断错误，对 9,000 多个账户，我们又抽取 1,000 个进行人工判断，根据该用户的回答问题列表做进一步判断，几乎全部账号都包含明显的广告推销嫌疑，因此本文假设上述 9,000 多账户均为虚假账号。

2.3　本章小结

为分析 CQA 系统，本章提出面向社区问答系统高效率爬虫技术，详细介绍爬虫系统的设计架构，爬取策略和执行方案，给出了 CQA 数据集的收集过程。本章对比收集数据和实际百度数据的规模，并从用户分类和问答数两方面分析了爬取抽样的合理性和均匀性。为后续分析，本章同时爬取了著名虚假交易网站“猪八戒”网中有关“百度知道”的虚假信息，抽取“百度知道”中的虚假账号。

第三章　CQA 系统中的用户行为

不同的 CQA 系统具有类似的组成主体和主体属性，如表1.1所示。考虑到“百度知道”依赖于百度搜索，成长较快，有最大的用户群和搜索价值，因此本文选用“百度知道”作为 CQA 系统的代表。本章重点对“百度知道”数据集进行统计分析，以用户和问题为分析单元，从正常账户和虚假账户两方面描绘 CQA 账户特性，分别展现两类用户和问题的属性差异。针对这些属性的分析是很有实际应用价值的，不仅可以描述 CQA 用户的行为规律，而且能够用于虚假账户鉴别。

为分析 CQA 系统用户的个体属性和问答行为规律，有必要对上章爬取的数据进行预处理。从数据库中分别选取 9000 个正常账户和虚假账户，抽取上述账户的个体属性信息，提问的问题信息，提问的问题的回答信息，用户回答问题的信息。选取账户数据集时发现，匿名提问和匿名回答的情况比较常见，且数量十分庞大，为避免影响对账户的特征分析，本文删除了以“jbp”开头作为匿名账户标识的数据，以“zdsjwy”开头的匿名手机知道账户数据，以及“知道贡献者”作为标识的匿名回答者。

3.1　用户的个体属性分析

“百度知道”的账户拥有提问数、回答数、经验值等多种基本属性。这些属于账户的个体属性，因为这些统计值描述了账户个体的本身性质，和交互的对象并无关联。“百度知道”给出了账户的成长体系和积分获取办法 [43] 。在“百度知道“系统中，回答数最多的 20% 的用户和提问数最多的 20% 的用户，分别贡献了对应的 81.9% 的回答和 76.3% 的提问。在这两类用户中，有大约 50% 的用户在提问和回答方面都有突出的贡献，成为建设知识共享平台的中坚力量。该结论符合众所周知的二八定律 [44]。文献 [42] 表明在视频共享的 Youtube 网络中，同样存在少数人贡献大多数资源的现象。用户的其他属性如

经验值、财富值和赞同数近似于幂律分布，即较少的用户拥有大数目的财富值，这也反映了大多数资源掌握在少数人手里的规律。

考虑到账户操作会对至少两个基本属性产生影响，因此我们计算各个属性之间的 Pearson 相关系数 [45]。该系数能够衡量任意两个属性间是否存在线性关系，系数值接近 1 或 -1 时，说明两属性间具有线性相关。本文发现，回答数和提问数相关系数为 0.01，二者几乎没有线性关系，但是经验值与财富值相关系数为 0.83，这可能是由于获取财富值的同时，通过“登陆”、“回答”等操作，一般都相应增加经验值。

基于账户的基本属性，我们提炼出新的具有物理含义的属性，来描述用户的个体属性。结合第二章给出了虚假账户的定义和样本，本节分析正常和虚假两类账户的不同属性。

3.1.1 用户回报率

本文定义回报率为用户的财富值与回答数的比值。定义该属性有着实际意义，因为该属性描述了平均一个回答所赚取的财富。由“百度知道”的成长体系可知，只有优秀的回答（被提问者采纳或者被网友赞同推荐）可以产生财富值，因此回报率可以体现回答者的专业程度。图 3.1给出了两类账户回报率的累计分布情况。可以看出超过一半的账户的回报率集中在 8 分以下。虚假账户比正常账户的回报率偏高，这主要是因为虚假账户的问答模式为上文提过的“一问一答”模式：威客自己拥有的百度账号回答的答案总能被自己采纳为最佳答案。另外，还可以看出有 5% 的正常账户的回报率十分高，达到 60 以上，相比而言，账户一个回答被采纳为最佳答案可以获取 20 分财富，可以看出这类账户的回答会在被选为最佳答案的基础上，得到更多网友的“赞同”，这些账户可以看做是某些话题的“专家”用户。另一方面，虚假账户由于张贴的内容有限，难以得到广大网友的“赞同”，所以其回报率集中在 8 到 20 之间，没有比较出众的回报率。

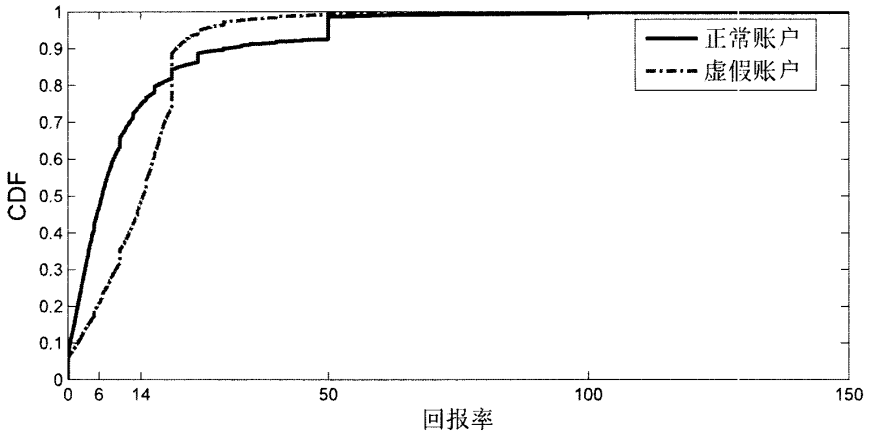

图 3.1 用户回报率分布

3.1.2 用户类别熵

“百度知道”问答平台包含了 14 个一级问题类别，图3.2给出了正常和虚假两类账户的提问活跃类别情况。横轴指提问类别活跃度，计算方法为：对两类用户提问的问题获取问题类别，计算不同类别的问题占据所有提问数的比例。在图3.2的问题类别中，正常用户活跃情况依次降低，而对于虚假用户，生活、地区、商业/理财类成为广告营销的主战场。

通常，一个用户在“百度知道”不止活跃在一个问题领域，因此考虑用户活跃的类别信息。假设共有 M 个问题类别 $\{c_1, c_2, ..., c_m\}$，单个用户的总提问数为 N 个，用户在每个问题类别的提问数为 $\{n_1, n_2, ..., n_m\}$, 收集用户提问的问题类别 $c_1, c_2, ...c_n$，定义问题类别熵 E_{cate} 如3.1所示。图 3.3 给出了两个用户的提问类别熵，根据“百度知道”提供的一级问题类别共 14 个，同理可以计算用户的回答类别熵。

$$E_{cate} = \sum_{i=1}^{M} -\frac{n_i}{N} \log \frac{n_i}{N} \tag{3.1}$$

对每个用户的活跃类别数进行统计，大部分用户活跃的一级分类数目小于4。如图3.3所示，正常用户的类别熵要小于虚假用户的类别熵，正常用户的提问类别熵平均值 0.18，虚假用户的类别熵平均值为 0.53，说明虚假用户会活跃

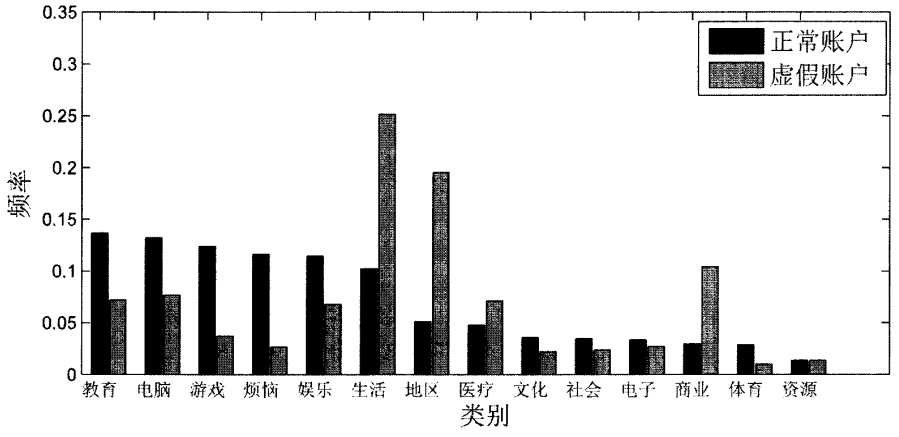

图 3.2 各类别的用户提问活跃度

在较多的问题领域，且倾向于没有轻重之分。这可能是因为虚假用户会因为接收任务多种而涉及到多种问题领域。同样用户的回答类别熵也表现出相同特性。如果将用户经常活跃的问题类型和类别熵结合起来，可有助于鉴别虚假用户。

图3.2中对于正常账户来说，排在前三位的问题类别属于客观知识类，相反，回答的活跃类别排在前两位的是烦恼和娱乐休闲，较多的属于主观知识类，说明网民对于客观知识类的疑问较多，不过相对轻松和主观的问题领域能收到最多的回答。分析单个用户提问类别与回答类别的相似性，发现超过 70% 的用户提问的问题类别和回答的问题类别没有交集，而有交集的用户大部分活跃在流行度较高的类别，例如娱乐、教育和电脑。而资源共享和文化、医疗和社会民生相对比较冷门，一方面是因为问题可以有多重类别，但倾向于归入流行度较高的类别，另一方面，冷门的类别对回答者的知识要求相对较高。由于虚假用户更可能通过回答正常问题来伪装为正常账号，减少被“百度知道”封号的风险，所以其回答类别倾向与正常账户差异较小，不过两类用户的提问类别倾向存在较大差异。其中“提问类别倾向”在下文的虚假账户鉴别中表现出较高的区分度。

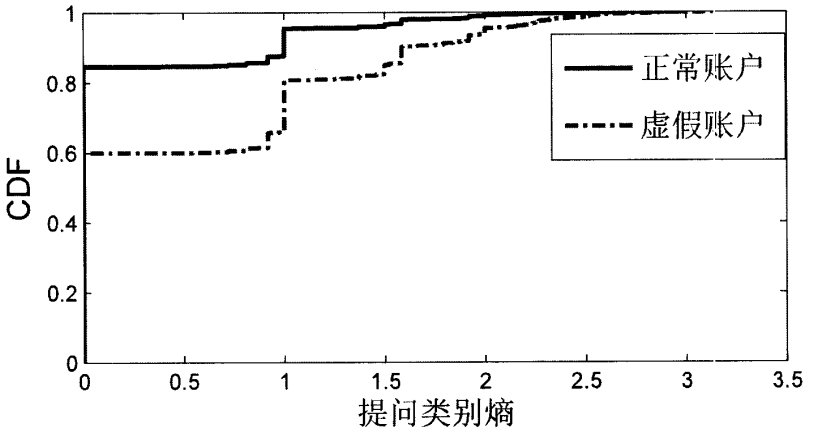

图 3.3 用户的提问类别熵

3.2 用户的问答行为研究

上节中分析了用户的个体属性，包括回答数，提问数等，因为这些统计值描述了账户个体的本身性质，和交互的对象并无关联。本节提出并分析了用户的行为属性，例如问答时间差，经验值的回答提问比等。这些属性描述了用户间相互关联，刻画用户的问答行为，包括用户参与的问题特征，问题中参与的提问者回答者属性间的关系等。

3.2.1 用户参与的问题属性

本小节介绍表1.1中“问题”的各属性间明显的影响关系，其中正常账户提问的问题统称为“正常问题”，虚假账户提出的问题统称为“虚假问题”。

问题解决率：用户提出的全部问题中被解决的问题所占的比例。由上节的用户数据集抽取的正常用户问题 5,818 个，虚假用户问题 13,971 个。如果问题有最佳答案就认为该问题已解决。这些问题中，虚假用户的提问解决率为 83%，远远高于正常用户 70% 的提问解决率。一方面是因为绝大部分虚假问题属于“一问一答”模式，另一方面虚假问题的悬赏分普遍偏高，这也吸引了不少的用户给出优秀回答。

问题收到的答案数：百度知道的单个问题平均有 5 个回答，所有回答中最多有 20% 的回答为满意答案，然而部分提问者并不设置满意答案，所以只有

13.8% 的回答为满意答案。文献 [4] 中通过人工审核的方法证实问题提问者选出的满意答案的质量的确要高于其他答案。实际上，本文分析数据发现，满意答案的回答者的特征都要优于其他回答者，比如最佳答案回答者的平均财富值为 10,870，非最佳答案回答者的平均财富值则为 4,067，满意答案的回答者的平均采纳率为 27%，而其他回答者则不足 15%。分析问题的回答总数，由图3.4 可知，超过 50% 的虚假问题最多有一个回答，而只有 20% 的正常用户最多有一个回答，而且虚假问题收到的赞同数小于正常问题。正常问题由于具有实际的参考作用和问题意义，所以比虚假问题受到更多人的关注。

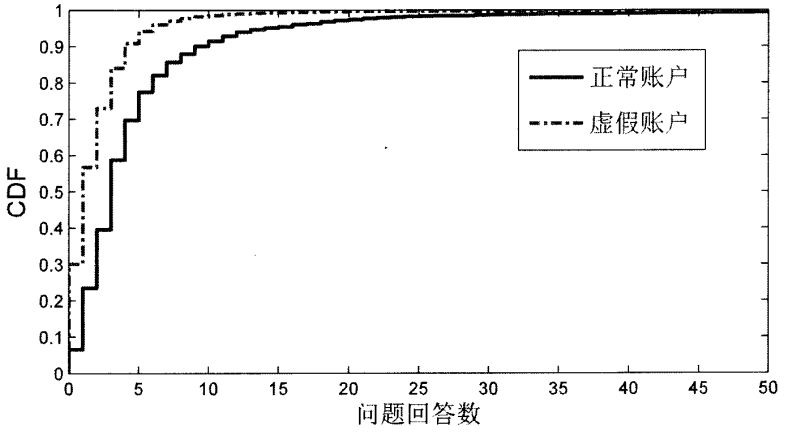

图 3.4 问题答案数分布

问题的浏览数：CQA 系统的知识库不仅能解决提问者的问题，也能够被互联网用户通过搜索引擎接触到，问题页面的浏览次数记录了该页面受到访问请求的次数，体现了该问题的参考价值。浏览次数越大，说明该问题越受到关注，其中的答案更有可能帮助到更多人。一般内容类似的问题，“百度知道”会将具有优秀回答的问题页面排在搜索结果前列，这样又增加了浏览可能性。图 3.5展示了正常和虚假两类问题浏览次数的累计分布，可以看出虚假问题的浏览次数明显偏低，超过一半的浏览次数不超过 40，而正常问题的浏览次数在 500 到 1,500 之间。虚假问题由于问题自身存在垃圾信息，搜索引擎会索引虚假问题并推送给网民，这样会影响到品牌搜索排名和用户使用体验。

问题的提问时刻：用户在单个问题中进行的问答交互行为，不仅包括提问者和回答者自身的基本属性特征，还包括提问者和回答者的行为时间。如图3.6

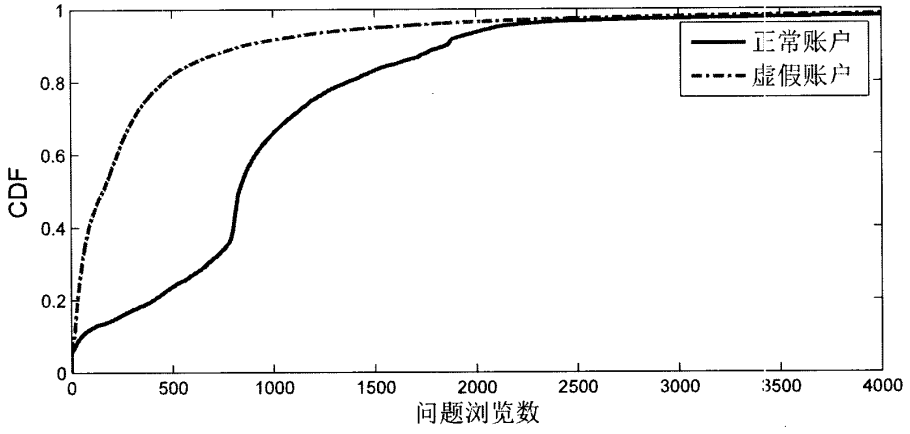

图 3.5 问题浏览数分布

给出了正常和虚假两类问题的提问时刻分布，横坐标为一天 24 个小时时刻，纵坐标为在对应时刻提问的问题频率。可以看出正常和虚假两类账户基本都符合网民的生活作息，但是明显地，虚假用户在 10 到 14 点的工作时间有突出的提问频率，正常用户在中午休息时间和晚上休息时间提问频率较高。这可能是由于虚假账户提问行为具有目的性和专业性，所以占用日常工作时间较多。

3.2.2 问题的回答者属性

本小节主要分析问题答案和回答者账户的个体属性特征：如表1.1所示的“回答”属性和“用户”属性。其中由3.1.2 可知提问行为更能体现正常账户和虚假账户间的问答差异，而本节将正常账户提问的问题统称为“正常问题”，虚假账户提出的问题统称为“虚假问题”。

问题经验值的回答提问比：定义为一个问题中最佳回答者的经验值与问题提问者的经验值的比值。图3.7 中展示的是问题回答者与提问者的经验值差距。横坐标表示问题经验值的回答提问比，纵坐标为比值的累计分布。可以看出正常问题中，回答者的经验值远远高于提问者经验值，但是对于虚假问题，超过 90% 的问题回答者与提问者的差距并不明显，这也侧面反映了虚假问题中的提问者和回答者是由同一个自然人伪造的，因此两类账户在“百度知道”上发展比较均衡。同样我们发现回答者与提问者的财富值比值，其分布也具有相同的性质。

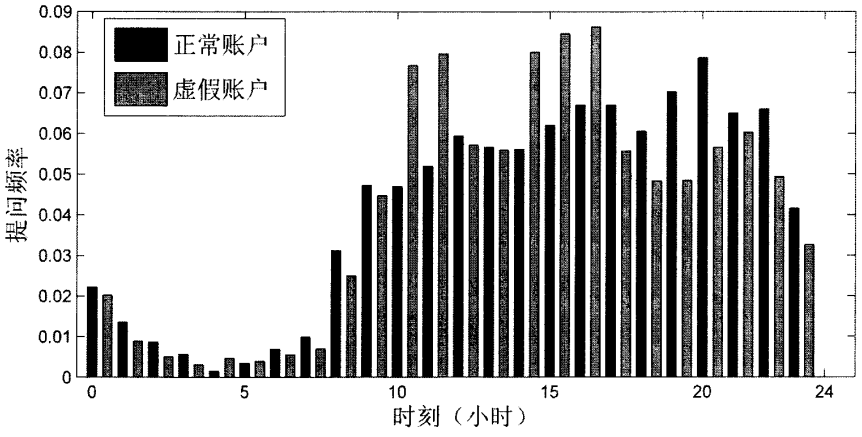

图 3.6 单日不同时刻的提问频率

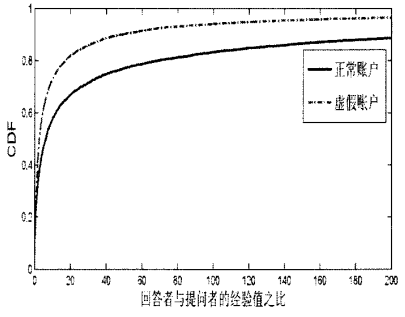

图 3.7 问题经验值的回答提问比分布

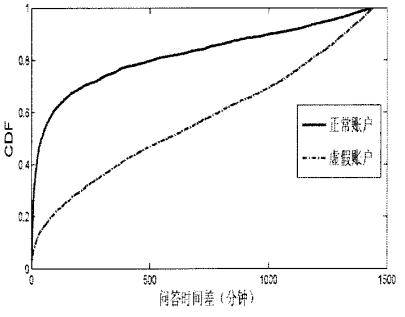

图 3.8 问答时间差的分布

问题的问答时间差：本文定义一个问题的问答时间差为最佳回答的张贴时间与提问时间之间的时间差。如图3.8，横坐标是问答时间差，以分钟为最小单位。最佳答案的张贴一方面表现了问题对账户的吸引力，一方面表明了提问者对答案质量的选择行为。不难看出超过一半的正常问题的时间差小于 30 分钟，远远小于虚假问题的时间差。这可能是因为正常用户在提出问题之后会在线关注该问题的解答状况，并倾向于选择最快回复的答案作为最佳答案。而作为虚假用户没有在线等待解答的需求，而且绝大部分最佳回答也是由威客自身提供，所以时间差较大。由上节可知“一问一答模式”占据了 90% 以上的任务，而模式中会特别规定回答者何时给出最佳答案，因此两类问题的问答时间差区分比较明显。下文在虚假账户鉴别中会基于问题的问答时间差计算账户的“问答时间差”属性，该属性在账户分类中起到重要的作用。

3.3 本章小结

本章揭示了问答系统中用户和问题两大主体的属性特征，对于每一个组成主体，本章又分别从正常账户和虚假账户展示行为规律的差异，包括两类用户的提问类别活跃度，问题解决率，答案数和浏览数的分布等。并进一步对行为差异给出合理的解释。此外，本章还提出一组具有物理含义的账户属性，包括“回报率”，“类别熵”，“问答时间差”等，这些属性在下文的鉴别虚假账户时具有较高的区分度。

此页不缺内容

第四章　CQA 网络建模及社团发现技术

第三章分析了 CQA 系统账户的个体属性和问答行为，然而账户个体性质不足以表现出 CQA 账户在全局系统中的问答交互性质，例如账户间最短距离，账户社团结构等。为分析账户间的社交关系，在本章中，我们使用社会网络分析方法研究 CQA 系统的账户网络，并提出一种基于标签传播的社团发现算法，实验证明该算法能够避免发现大量无效的微型社团，并且有效检测出 CQA 系统中有意义的社团。

4.1　CQA 系统中的账户网络建模

由第一章定义，“CQA 账户网络”是指 CQA 账户通过问答行为形成连接关系而构建的网络模型。本节提出了 CQA 账户网络的两种构建模型：问答完全网络和回答环形网络。两种网络模型有较强的连接性，容易传递信息。应用“百度知道”数据，本文构建 CQA 账户网络并给出 CQA 账户网络的各项属性值，分析表明 CQA 账户网络是典型的复杂网络。

4.1.1　建立 CQA 账户网络模型

本文在基于有向图建立网络模型，并进行社团发现时，由于一个提问者由于对应很多回答者，所以提问者成为了小社团的核心，这导致检测到的社团质量较差。因此本文使用无向有权图模型来描述 CQA 系统中的账户网络。其中 CQA 系统中包含的 m 个用户形成节点集合 V，节点间出现的问答关系构成边集合 E。

定义 4.1.1. 问答完全网络：以用户为节点，问答关系为边。如果两个用户节点 v_i 和 v_j 在 k 个问题中同时出现（分别为提问者和回答者或者同为回答者），那么 v_i 和 v_j 之间存在一条权值为 k 的无向边 e。按照上述定义，同一问题的回答者在同一个问题页面给出回答或者互相参考评论，并为提问者提供帮助，这些用户形成一个小型的完全图，即如果该完全图中包含 n 个用户节点，那么会有 $n(n-1)/2$ 条边。

定义 4.1.2. 环形回答网络：如果 m 个用户 $v_1, v_2, ..., v_m$ 共同回答某个问题，其中 $v_1, v_2, ..., v_m$ 按照回答问题的时间顺序排序，使用 $m-1$ 条边将 $m-1$ 对节点 $((v_1, v_2), (v_2, v_3), ..., (v_m, v_1))$ 连接形成一个环，则形成环形回答网络。假设节点 v_i 与 v_j 在某个问题中分别回答 n_1, n_2 次，则连接 v_i 与 v_j 的边 e_{ij} 的权值为 $(n_1 + n_2)/2$。

上述两种网络模型能够反映 CQA 用户不同方面的行为特征。在完全网络中，包含了用户作为提问者和回答者两方面的信息。明显地，如果用户 v_i 参与到一个属于热点话题的问题或者时事热点中，例如第三章中提到的“烦恼”类，v_i 的节点度数会非常大。完全网络将用户的活跃水平和问题流行度结合起来，非常适合研究热点话题检测和关键用户行为分析。环形网络仅仅包含用户作为回答者角色的信息。在环形网络中，节点的度数仅仅由用户的回答数决定。环形网络反映用户的兴趣以及具有相似回答行为的社群特征，具有共同知识领域的用户间关系。由于“百度知道”中的用户扮演两个角色“提问者”和“回答者”，而回答环形网络模型将提问者的角色从用户中剥离出去。对于类似 CQA 系统以内容为主导的社交媒体，有必要分析内容资源拥有者的网络特征。

本文使用以下指标来衡量一个网络的特性：幂率指数 [46]，连通比例，集聚系数 [46]，平均测地线长 [46]，同向匹配系数 r [47]，内部连接序列和社团结构，并选取三组数据来测试网络模型的可用性。许多真实的网络的节点度分布服从幂律分布：$p(k) \sim k^{-\gamma}$ [1, 9, 46]，其中文献 [48] 验证了几种抽样方法得到的抽样网络都会保留幂律分布特性，其中 γ 称为幂律指数。连通比例是指网络中的最大连通分量所包含的节点占网络总节点数的比例，用于衡量整体连通性。集聚系数描述网络中的节点之间结集成团的程度，一个节点的集聚系数是节点邻居之间的连接数与邻居之间最多的连接数的比值。而网络的集聚系数 CC 是所有节点集聚系数的平均值，比随机网络高的聚集系数意味着节点有聚集倾向。测地线 G 是网络中两个点之间的经过的最少节点数的路线，其中经过节点数即为测地线长。网络的平均测地线长是可到达的节点对的测地线长的平均值。同向匹配系数 r 测量的是互相连接的节点有权度的相关性，通过节点对之间的 Pearson 相关系数衡量。正值的能够测量具有类似度数的节点之间的相关性。内部连接序列 CS 定义如下：如果按照节点的有权度数从小到大排序，并将节点切分为 N 份，则第 i 份节点集合中不同节点之间的连接总数记为 l_i，而 $\{l_1, l_2, ..., l_N\}$ 记为网络的内部连接序列。社团结构反映了用户的交互特征和群

聚性质。

4.1.2　CQA 账户网络属性分析

本文从娱乐、生活和资源三种类别中分别挑选 400,000 个问答，使用这些问答中的参与者作为节点，两个节点的邻接由问答关系决定。这三种类别流行度依次降低。将每组数据分别构建问答完全网络和回答环形网络可以得到六个具体的 CQA 账户网络。表4.1和表4.2 给出了六个网络的各项指标值，从表中我们可以得出如下结论：

表 4.1 问答完全网络的属性值

类别	顶点	边	幂指数	连通分量	聚集系数	同配系数	平均距离
娱乐	265458	1246790	1.88	74.979%	0.44504819	0.136297	10.93
生活	289062	1119264	1.84	69.902%	0.37828876	0.165934	10.08
资源	177961	421587	1.99	38.612%	0.60050119	0.46295	12.43

表 4.2 回答环形网络的属性值

类别	顶点	边	幂指数	连通分量	聚集系数	同配系数	平均距离
娱乐	206523	256487	2.58	70.7%	0.03345036	0.110111	6.72
生活	216458	268678	2.5	63.413%	0.03649175	0.148347	6.45
资源	123448	118781	3.2	33.933%	0.09475506	0.165241	8.69

- 无标度性质：无标度网络中的大部分节点只和很少节点连接，而有极少的节点与非常多的节点连接，属于复杂网络的一种网络类型。现实中的许多网络都带有无标度的特性，例如因特网、金融系统网络、社会人际网络等等。表 4.1 和4.2给出了六个网络的节点分布的幂律指数，如图 4.1中的生活类网络所示，横坐标为节点度数，纵坐标为互补累计分布，两种基于抽样数据的网络模型均体现了原始真实网络的幂律特性，但是 γ 数值不同。可以看出大多数节点度数偏小，只有少部分节点拥有相当大的度数，本文称这类度数较大的节点为枢纽节点。考虑到问答页面新产生的速度远远大于问答页面一般被“百度知道”删除的速度，随着时间的发展，“百度知

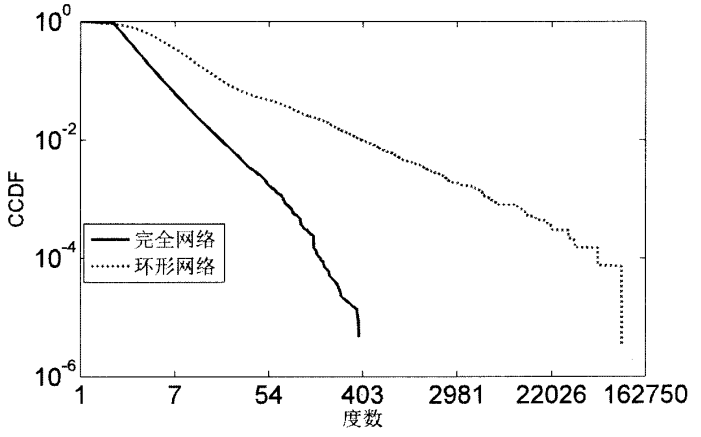

图 4.1 生活类网络节点度数的 Log-Log 分布

道”系统进一步扩大，网络中的平均度会逐步增加，且部分节点的度数可以变得十分巨大。对于 CQA 网络，如果随机移除部分用户，网络中剩余的大部分用户依旧可以连通，但是如果连续移除高度数的节点，网络的连通性会迅速降低。对于生活类完全网络，当移除 3% 的高度数节点时，连通比例下降为 30%。无标度网络能够提高网络抗破坏的鲁棒性，可以承受强大的意外故障，但面对协同性攻击时候比较脆弱，所以枢纽用户为 CQA 系统做出相当大的贡献，但又容易成为协同性攻击的目标。

- 富人俱乐部性质：网络中的枢纽节点内部相互连接紧密称为富人俱乐部性质。以生活网络为例研究内部连接序列，在图4.2中，将生活类网络中的节点根据节点度数大小分为 40 份，横轴表示生活类的内部连接序列，纵轴表示每份节点集合的内部连接数。可以看出完全网络中随着节点度数的增加，内部连接序列 CS 的序号增加，各序列的内部连接数快速增加，度数最大的节点集合之间连接数 l_{40} 远远大于其他节点内部连接数目；当从完全网络中除去越来越多的低度数的节点时，可以发现网络的聚集系数快速增加，这些都说明网络中大度数节点内部连接紧密的性质，即存在富人俱乐部现象：拥有大量边的少数节点更倾向于内部互相连接。在下文的社团检测结果中我们发现，度数较大的前 2.5% 节点所在的社团重叠性较强，进一步说明大度数节点之间联系紧密，形成一个或多个重叠的“俱乐部”。

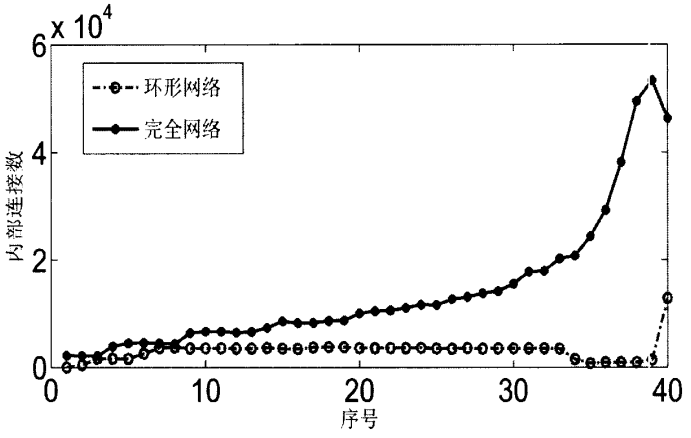

图 4.2 生活类网络的内部连接序列

进一步分析第 40 份的大度数节点参与的问答情况，其中回答提问经验值比小于网络中的平均值，可以看出在 CQA 系统中，这些枢纽节点更像是在讨论问题，而不是仅仅获取问题答案。

- 小世界性质：当网络中大部分节点从任一节点经过几步就可以达到，则称该网络具有小世界性质。表 4.1 和4.2的平均地测距离均小于 13。即一个用户平均至多路过 13 个节点就可以和目标节点连接。生活类和娱乐休闲类的两种网络中大多数节点对都是可以连通的，资源共享类别的两种网络具有较低的连通比例，这可能和该类别较低的活跃度以及资源类清晰的专业边界相关。观察图4.2中环形网络的内部连接序列，和完全连接网络的连续增长情况不同，内部连接数呈现上升、稳定、下降再突然上升趋势。实验发现不论内部连接序列分为多少等份（10-40），在 85%-95% 之间的节点连接数都会突然减小。我们称该部分具有最小内部连接数的节点为颈部节点。分析颈部节点的连接对象，发现大部分节点和更高度数的节点连接，一部分边与低度数节点连接。可以看出颈部节点起到一种承上启下的作用，虽然本身度数较大，但是大多数连接指向度数不相配的节点。娱乐类和资源共享类同样具有相同的内部连接性质，颈部节点为低度数用户连接到高度数用户提供关键路径的作用，对增加网络连通性，缩减网络中的平均地测距离起到了一定的作用，为小世界网络性质提供客观实体。

文献 [46] 给出了 30 种复杂网络的属性特征，网络类型覆盖了以“电影合作演员”，“邮件地址”等为例的社交网络，以“索引网络”，万维网为例的信息网络，以“因特网”，P2P 网络为例的技术网络，和以“神经网络”，“蛋白质互作用网络”为例的生物网络，属性特征涵盖了幂指数，聚集系数，同向匹配系数，平均地测线长等网络属性。将本章基于 CQA 系统数据构建的 6 个网络和文献 [46] 中列举的 30 个网络相比，可以发现 CQA 网络的性质与社会网络性质相似，特别是生物合作作者网络。CQA 网络和其他类型的网络有不匹配的数值，例如技术网络和生物网络。上述 6 个网络自身相比，具有较高流行度类别的 CQA 网络有更强的连接度。两种 CQA 网络模型均是合理的，比分散的用户个体属性反映出用户行为更多的性质。

4.2 CQA 网络中的社团发现

上节中给出了两种网络模型，并分析了两类网络的基本属性。本节首先提出一种社团发现算法 MSLPA（Multilayer speaker-listener label propagation algorithm), 该算法能够高效发现网络中合理的社团。基于该算法的发现结果，本文重点分析虚假账户，正常账户，以及颈部节点社团，枢纽节点社团的社团特征。

4.2.1 MSLPA 社团发现算法

由第三章可知大多数用户活跃在至少两个类别，“百度知道”中问题的类别以树型结构组织起来，所以用户社团必定是相互重叠的和具有层次性的。为了探索用户如何交互，我们提出一种 MSLPA（Multilayer Speaker-listener Label Propagation Algorithm）算法来检测有意义的社团。

文献 [38] 中提出的基于标签传播的 SLPA（Speaker-listener Label Propagation Algorithm）算法可用于重叠社团检测。SLPA 算法的时间复杂度和网络中的边数成线性关系，时间复杂度为 $O(tm)$，其中 t 表示标签传播过程的迭代次数，m 表示边数，因此适用于大规模的稀疏网络。然后实验中发现 SLPA 算法在应用于“百度知道”网络中时会检测出大量的微型社团，分析微型社团的成员属性，其中一方面是由于数据中存在一定量的的非活跃账户，因为交互性不

足无法形成有规模的社团；另一方面还有不少社团之间粘合度较高，且账户具有同质的行为，却无法形成一个社团整体，即 SLPA 算法不能够充分的整合同质的用户群组。

鉴于 SLPA 算法的不足，本文提出了 MSLPA 算法。该算法能够弥补 SLPA 算法的缺陷，充分融合小型社团来整合微型社团。与 SLPA 相比，MSLPA 在执行过程中采用新颖的方法衡量节点间的距离，进而更新节点和边权值，不断生成新的社团网络。而在每层社团网络中，借鉴 SLPA 的标签传播思想，凝聚网络节点。

MSLPA 的迭代过程由如下五步：标签传播，生成社团，构造新节点集合，构造新边集合，网络更新。在 $G=(V,E)$ 中，每个节点是仅具有一个账户成员 v_0 的初始社团 c^0。每个社团的“标签”是成员 v_0 的 ID。初始时以社团为节点形成的社团网络记为 $G^0=(C^0,E^0)=G$。在标签传播的迭代之后，MSLPA 生成了第一层社团集合 $C^1=\{c_1,c_2,\ldots,c_m\}$。分配新的社团 ID 给每一个新社团 c_i。如果 C^1 中的社团 c_i 和 c_j 包括 k 个共同成员节点，则使用一个权值为 k 的新边 e_{ij} 连接两个社团。社团之间边的集合记为 $E^1=\{e_{12},e_{13},\ldots,e_{mn}\}$。在网络 $G^1=(C^1,E^1)$ 经过同样的标签传播迭代过程之后，我们可以得到第二层社团 $C^2=\{c_1^2,c_2^2,\ldots,c_k^2\}$。相似的，使用相同的规则可以构建第二层社团网络 $G^2=(C^2,E^2)$。MSLPA 产生超过 min 个的社团。随着融合的次数增加，互相分离的社团对比例越来越大。当社团网络中的社团数量少于 min 时，MSLPA 停止。算法 4.1描述了整个 MSLPA 算法过程。MSLPA 的时间复杂度为 $O(h(tm+c^2)$，其中 t 表示标签传播过程的迭代次数，m 表示边数，c 为社团数，h 为网络中潜在的社团层次数。

为了衡量社团发现算法的性能和结果有效性，本文定义如下指标：兴趣匹配度 α 和同质社团比例 $Ratio$ 来分析社团发现算法。

假设一个账户回答某个问题，说明该账户对该问题所属的类别感兴趣，相对提问者来说更擅长该类别的知识内容。社团网络中的社团节点具有如下三个属性：社团规模 $CSize$，兴趣集合 $intC$，核心兴趣集合 $CoInt$。$CSize$ 是社团中账户个体成员的个数，初始社团网络 $G^0=(C^0,E^0)$ 中每个社团节点的 $CSize$ 则初始化为 1。如果账户 v 回答问题 q，q 的类别称为用户 v 的一个兴趣 int，属于账户的兴趣集合 $intC$。账户的问答信息可以初始化 C^0 中每个元素的 $intC$

```
Input: N(Nodes, Edges) is a network; Iter is the times of iteration; Min
       is the minimal number of members in one community; Thre is
       used to select labels.
Output: The detected communities
index = 0;Commu = {}; NewNodes={}; NewEdges={};
repeat
    repeat
        Random Sending-Order of Nodes;
        for each node β in Nodes by Sending-Order do
            the neighbors of β add Label_β into their own Lablelist;
        end
    until index ≥ Iter;
    for each node β in Nodes do
        Pick up Set α {label| propotion of label > Thre} from
        β's Label list, push α into Commu[label];
    end
    Construct Sets NewNodes{Commu|{Commu}| > Min} from Set
    Commu;
    for every node pair (Commu_i, Commu_j) in NewNodes do
        if |{Commu_i} ∩ {Commu_j}| == k then
            NewEdges .push( edge(i, j, k) );
        end
    end
    N ← NewN{NewNodes, NewEdges} ;
until |Nodes| ≤ Min;
```

算法 4.1: MSLPA

属性。*IntC* 由公式4.2 计算得到。社团网络 *CNet* 中，社团节点 C 包含成员 $c_1, c_2, ...c_m$，当成员 c' 中的 *intC* 中兴趣 *Int* 的比例大于阈值 λ 时，选取该兴趣 *Int* 作为社团节点 C 的 *Int*。即从成员的兴趣集合中选出代表性的兴趣组成社团的兴趣集合。

CoInt 可以由公式4.5计算得到。在社团节点 C 的兴趣集合 *IntC* 中，如果某个兴趣 *Int* 在兴趣集合中的比例大于阈值 θ，则该兴趣称为 C 的核心兴趣。如果 C 的核心兴趣集合 *CoInt* 中存在至少一个兴趣 *Int*，意味着社团 C 把超过一定比例的相似子社团融合在一起，该社团称为一个有意义的同质的社团 *HomoC*。值得强调的是 *CoInt* 和 *intC* 意义并不相同。计算 *intC* 的 λ 和社团

C 成员的所有兴趣相关，而 θ 仅和社团本身兴趣相关，$intC$ 连接底层社团成员与当前社团，$CoInt$ 连接当前社团与未来将要生成的大社团。为了探索社团的演化过程，MSLPA 能够记录每一层社团和 ID 之间的映射关系，以及所有社团的成员组成。对于第 n 层中的社团集合，可以追溯第 i 层的社团聚集过程，其中 $i \leq n$。

兴趣匹配度 α：“百度知道”账户在填写个人资料时可以标记自己的“擅长领域”，用以表明自身的专业领域。为了检测社团是否合理和有意义，本文使用账户标记的擅长领域作为社团主题的基准兴趣，考察算法发现的社团核心兴趣 $CoInt$ 是否和社团成员用户的真实兴趣相匹配。由于部分用户未标记自身的“擅长领域”，所以本文只考虑超过 1/4 的成员标记“擅长领域”的社团，符合条件的社团约占整个社团集合的 2/3。定义兴趣匹配度 α 如4.1，其中 $CoInt$ 为社团网络 $CNet$ 社团的核心兴趣，$ExpInt$ 为按照获取 $CoInt$ 的方法选取的自标记兴趣，其中账户 v 的 $intC$ 不是由问答关系决定，而是由用户自身标记的“擅长领域”决定，$CoInt$ 与 $ExpInt$ 具有相同的类别越多，α 越大；兴趣匹配度 α 衡量了“百度知道”账户自己标记的擅长领域与算法检测出的社团核心兴趣的一致性。由于账户在设置“擅长领域”时可以选择类别的任一层次，所以在计算 α 时，当 $CoInt$ 与 $ExpInt$ 之间互为上下级类别或同属于同已上级类别时，$|CoInt(C) \cap ExpInt(C)|$ 可以适当地赋予一定的正值。

$$\alpha = \sum_{C \in CNet} \frac{|CoInt(C) \cap ExpInt(C)|}{|CNet| \times min\{|CoInt(C)|, |ExpInt(C)|\}}, \tag{4.1}$$

同质社团比例 $Ratio$：本文定义公式4.4中的 $Ratio$ 来衡量社团发现算法的结果。θ 越大，说明对同质社团的要求越严格；$Ratio$ 越大，意味着社团网络包含了更多的同质社团，也即发现算法生成了质量更好的社团。

$$IntC = \{Int^* | \frac{\sum_{c' \in C} \sum_{Int_i \in IntC(c')} \phi(Int_i = Int^*)}{\sum_{c' \in C} |IntC(c')|} > \lambda\}, \tag{4.2}$$

其中 $\lambda \in (0,1)$ 且

$$\phi(Int_i = Int^*) = \begin{cases} 0, & if \quad Int_i \neq Int^* \\ 1, & if \quad Int_i \equiv Int^* \end{cases} \tag{4.3}$$

$$\begin{aligned} Ratio &= |\{HomoC\}|/|CNet| \\ &= \frac{|\{C \mid C \in CNet \text{ and } CoInt(C) \neq \emptyset\}|}{|CNet|}, \end{aligned} \tag{4.4}$$

其中

$$CoInt = \{Int^* | \frac{\sum_{Int_j \in IntC(C)} \phi(Int_j = Int^*)}{|IntC(C)|} > \theta\}. \tag{4.5}$$

由于 CQA 系统是以内容为中心的社交媒体，因此社团的主题至少有一个方面是和兴趣相关，下一节在“百度知道”数据集中应用多种社团发现算法来检测社团，并从社团兴趣等方面比较算法结果。

4.2.2 社团发现算法性能评估及结果

本文以 9,000 个正常账户和 9,000 个虚假账户为种子，采用 snowball-sampling 的方式从数据库中获取实验数据，方法如下：首先获取上述账户的回答问题列表，扩展账户之间的边；其次添加上述问题的回答者账户，扩充网络中的节点；再获取上述账号的回答问题列表，依次迭代进行。基于上述账户节点和问答边，使用完全网络和回答环形网络构建两个具体的“百度知道”网络，其中环形回答网络共包括 141,043 个节点，140,771 条有权边；完全回答网络共包括 170,456 个账户节点，464,054 条有权边。除 SLPA 与 MSLPA 算法之外，本文还使用下述三种算法发现社团，与 MSLPA 算法做对比。

CPM（clique percolation method）算法 [32] 假设一个重叠社团是由完全连接的子图构成，通过搜索邻接 cliques 来发现社团，其中一个 k-clique 是指由 k 个节点组成的一个完全子图，而一个 k-clique 社团是指由多个 k-clique 连接成的集合。由于类 CPM 算法旨在发现网络中特定的局部结构，所以更像是模式匹配，而不是发现社团。Cfinder[49] 是 CPM 的一个算法实现，该程序可以发现重叠类的有权重的社团，时间复杂度和网络结构相关。OSLOM 算法 [39] 是将网络中的一个聚类与一个随机网络对比统计显著性，根据显著性结果生成社团。假设某个已有社团的邻居节点 v, 把 v 添加到社团中产生比添加到随机网络中可以产生更多内部连接数的累积概率记为 γ，当最小的 γ 的累积分布概率小于某个阈值时，算法认为 v 是显著的，可以添加 v 到相应的社团中。其中 [50] 是该算法的一个实现，能够发现重叠类、层次性的有权重的社团，时间复杂度

为 $O(n^2)$。Link[51] 算法则是把社团看做是连接的集合，而不是账户的集合，如果网络中和一个顶点相连的边都被分到同一个聚类里时，称该节点是重叠的。连接通过边相似性的层次性聚类划分，当两条边的相似度小于某个阈值时，就切断两个边的关系。[52] 是该算法的一个实现，能够检测出重叠性的无权重类的社团，时间复杂度为 $O(nk_{max}{}^2)$，其中 k_{max} 表示最大的节点度数。

1. 不同社团发现算法在两种 CQA 账户网络中的兴趣匹配度 α 比较：

表4.3和表4.4给出了五种算法的社团检测结果，其中 t 表示迭代次数，m 表示边数，n 表示节点数，c 为社团数，“-” 表示无法给出结果。考虑到社团的有意义性，本文令社团的 $CSize$ 至少为 3，所以生成的社团只包含一定比例的账户节点。

表 4.3 回答环形网络中社团发现算法的性能比较

算法	运行时间	社团总数	兴趣匹配度	包含节点比例	层次性检测
MSLPA	1200s	17050/9739/7080	40%	80%	可以
SLPA	60s	17050	40%	93%	不可以
Cfinder	50s	964	47%	10%	不可以
OSLOM	1200s	480/163	46%	10%	可以
Link	310s	20511	8%	94%	不可以

表 4.4 问答完全网络中社团发现算法的性能比较

算法	运行时间	社团总数	兴趣匹配度	包含节点比例	层次性检测
MSLPA	1381s	13450/11289/10743	45%	80%	可以
SLPA	50s	13450	45%	88%	不可以
Cfinde	-	-	-	-	不可以
OSLOM	3650s	110290/40256/7814	33%	- 97%	可以
Link	-	-	-	-	不可以

从表4.3可以看出，对于比较稀疏的回答环形网络来说，所有算法均能较快地给出社团结果，其中 MSLPA 耗用了相对较多但是能够忍受的时间。当 $CoInt$ 与 $ExpInt$ 都只包含一个元素时，除了 Link 算法外，其他四种算法的兴趣匹配度都稳定在 40% 以上，考虑到用户不止擅长单个领域，所以当 $ExpInt$ 扩展为两个元素时，兴趣匹配度提高到 70%；尽管 CFinder 和 OSLOM 的兴趣

匹配度偏高，然而发现的社团包含极少比例的节点数，这样会排除掉部分兴趣不明显或者不匹配的社团；相比较，MSLPA 则在考虑了大多数节点的情况下依然保持良好的兴趣匹配度，MSLPA 总体表现良好。

从表4.4可以看出，在比较稠密的问答完全网络中，CFinder 和 Link 无法处理数十万条边的网络，但 MSLPA，SLPA 和 OSLOM 运行时间变化不大，由此可以看出 MSLPA 在对网络的可扩展上表现良好。完全网络相对环形网络边数增加几倍，MSLPA 检测出更多的社团，OSLOM 首先生产了大量的迷你社团，但在后续的层次聚合过程中逐渐融合社团，SLPA 则不能够给出层次关系。整体来看，MSLPA 与 OSLOM 更适合于 CQA 网络中社团检测，同时还能给出社团的层次结构，前者在稀疏性网络或者网络结构简单的情况下表现优异，后者随着网络中边的增加，聚合能力要优于 MSLPA。

综合上述结论，相比 SLPA，MSLPA 添加了发现社团间层次关系的功能，使用社团间的重叠关系构建新颖的“社团网络”，同时效果上避免了产生大量的迷你社团，加强了社团间的聚集力度。相比 OSLOM，MSLPA 在常见的稀疏网络中表现良好且稳定。SLPA 适应于大规模社会网络中的社团检测，但不能检测社团层次性；CFinder 和 Link 仅仅适合具有特定网络结构中的社团检测。MSLPA 在稠密网络和稀疏网络中均有不错的表现，而 OSLOM 更适合在稠密网络中发现社团。考虑到算法中“标签传播”的基因，MSLPA 在完全网络中的社区发现结果差强人意，可能是因为完全网络中与一个问题相关的各节点之间具有同样的连接数，且该连接数和该问题的受关注程度相关，使得用户之间关系受到话题内容的干扰。

2. MSLPA 算法在环形回答网络中的同质社团比例 *Ratio* 比较：

在对比了几种算法的发现结果之后，为重点分析 MSLPA 的层次聚合效果，本文选取回答环形网络模型中发现社团，回答环形网络中用户角色均为回答者，可以考察一个社团如何对 CQA 网络知识库做出贡献。我们使用 snowball-sampling 方法选择两个数据集，第一个数据集 *Life_Net* 包含了生活类别下的问答记录，第二个数据集 *Mix_Net* 包含所有类别下的问答记录，两个数据集代表了不同层次的 CQA 网络。考虑到 MSLPA 在环形回答网络和完全问答网络中的效果差异不大，所以本文仅分析对环形回答网络。数据集基本情况如表4.5所示。

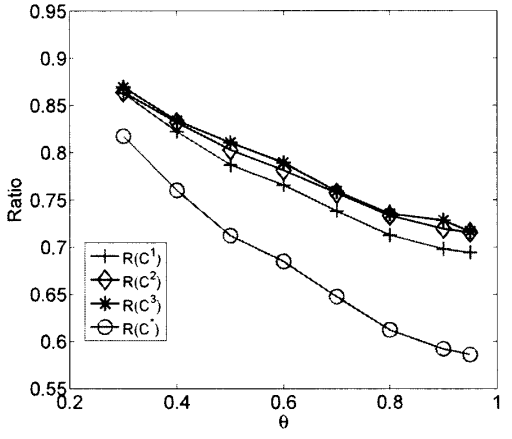

图 4.3 *Life_Nets* 中社团的 *Ratio*

图4.3给出了 *Life_Net* 发现过程中不同层次的 *ratio*。对于公式4.5的每一个阈值 θ，*ratio* 会随着社团网络的层次增高而增加。这意味着 MSLPA 在建立网络的过程中将兴趣相同或相似的社团聚合在一起。当 θ 大于 70%，*ratio* 超过了 75%；而在完全网络中，当 θ 设为大于 90% 时，*ratio* 超过 80%，由此也可以看出形成一个社团的主要因素是用户的兴趣。

和 SLPA 相比，MSLPA 能够避免微型社团的形成。实际上 MSLPA 生成的第一层社团网络等价于 SLPA 的结果。但是 MSLPA 中网络更新的过程将微小社团依附到相关社团中，从而聚合了相关社团。表4.5 给出了每层检测到的社团集合。MSLPA 减少了 20% 的社团数。对于每一种网络，网络所处的层次越高，平均社团大小就越大。在图4.4中，$c^{1.5}$ 是 c^1 中互相结合形成 c^2 的社团子集，而 $c^{1.5}$ 的平均社团大小远远小于 c^1 的社团大小。所以 $c^{1.5}$ 是 c^1 中的微小社团群。和 SLPA 相比，MSLPA 可以充分的融合社团，特别是那些微小社团。MSLPA 中的网络更新过程是融合社团的主要动力，大多数 CQA 网络可以最终稳定成一个社团集合，即不再出现融合的过程。结果显示具有更小 *ratio* 和更小 *CSize* 的微型社团能够被 MSLPA 融合。所以在社团层次结构发现以及微型社团聚合方面，MSLPA 比 SLPA 具有更好的效果。

MSLPA 不仅能够得到更合理的社团，而且在具有层次性结构的网络中表现良好。*Mix_Net* 包含从 14 个高级类别中选出的 Q&A 记录，在图 4.4 中，C^1, C^2, C^3 和 $C^{1'}, C^{2'}, C^{3'}$ 是 MSLPA 在 *Mix_Net* 社团网络中发现的社团集

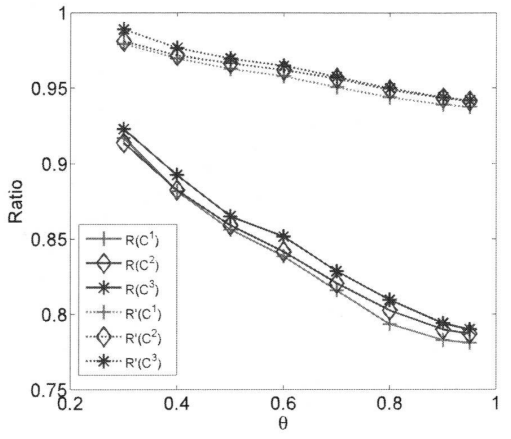

图 4.4 *Mix_Nets* 中社团的 *Ratio*

表 4.5 MSLPA 算法在不同层次结构网络中的融合效果

类型	属性	G^0	G^1	G^2	G^3
Life_Net	社团数	400,000	23638	16531	15992
	CSize	1	8.3	10.9	11.4
Mix_Net	社团数	73080	10409	8818	8312
	CSize	1	6.9	8.2	8.7

合，按照底层小类别计算前者的同质社团 *Ratio*，按照高层类别计算后者的同质社团比例 *Ratio′*。可以看出 *Ratio′* 远远大于 *Ratio*，意味着具有不同的底层类别但具有同样的高层类别的用户，更倾向于被融合到同一个高层类别中。虽然 MSLPA 没有和兴趣（类别）相关的先验知识，它仍然能够把同属于同一高层社团的子社团融合在一起。

4.3 本章小结

本章使用社会网络分析方法探索 CQA 系统的网络性质，给出了两种网络模型：完全问答网络与环形回答网络。分析基本网络属性可知 CQA 网络具有较强的社会网络性质，包括无标度性质，富人俱乐部性质和小世界性质。网络属性分析表明 CQA 系统中存在以用户兴趣为主题的账户社团。为分析 CQA 账户网络的账户聚集行为和社团结构，本章提出了面向 CQA 系统的 MSLPA 社团发现算法。将社团主题与账户自标记的主题进行对比，本章结果表明 MSLPA 算法能够有效发现有意义的、重叠的、具有层次结构的账户社团。将该算法与

已有的四种社团发现算法进行比较，实验结果表明 MSLPA 算法添加了发现社团间层次关系的功能，使用社团间的重叠关系构建新颖的“社团网络”，同时效果上避免了产生大量的迷你社团，加强了社团间的聚集力度，最适应于 CQA 系统的大规模稀疏网络。

此页不缺内容

第五章　CQA 系统中虚假账户的鉴别方法

以“百度知道”，“Yahoo! Answer”等为代表的一类 CQA 系统可以被搜索引擎爬取，未注册的用户也可以接触到其知识库信息，因此这种开放性的 CQA 系统特别容易受到恶意账户或者水军的入侵。本章节集合上文的分析结果，包括第三章对两类账户的属性分析，第四章中的社团发现技术，将上述分析结论和技术服务于虚假账户鉴别方案中，提高鉴别的准确性。

5.1　基于分类器的虚假账户鉴别

百度知道协议根据张贴的内容信息分为四大类，其中“广告营销类”成为最难鉴别和阻止的一类虚假信息。由第二章定义可知，本文将 CQA 系统中的账户根据发帖意图分为两类：正常账户和虚假账户。其中正常账户包括确实有需求而发布问题的提问者，和确实愿意提供与问题相符的答案的回答者；虚假账户指用户在发帖中另有所图，意欲欺诈的提问者和回答者，在本文中特指在 CQA 系统中发布广告营销信息的专用账户，例如从猪八戒网上获取的相关 CQA 系统账户。

为抵制垃圾信息对 CQA 系统的侵入，急需面向 CQA 系统的虚假账户鉴别方案。虚假账户鉴别一般使用分类技术，通过建立成熟的特征空间或者应用有效的分类器获得更好的分类性能 [23, 53]，文献 [54] 给出一种鉴别 twitter 上虚假账号的 $SSDM$ 算法，把微博账户之间的网络关系和内容整合在类 SVM 模型中，精度达到了 85%，相比而言基于内容的分类器和仅基于网络的分类器的精度只有 78%。然而该算法复杂度偏大，不适合实际环境中大规模数据。文献 [41] 中使用 SVM 分类器获得了 98% 以上的准确率，同时也给出了基于阈值的几条规则分类器，尽管该规则分类器十分简单，却可以获得和 SVM 类似的准确率，比较适用于大规模数据。在人人网等具有显式好友关系的在线社交媒体中，当一个陌生用户像一正常用户申请好友时，正常用户会判断陌生用户是否

有虚假用户的嫌疑，从而决定是否“接受好友申请”。人工判断结果体现在了一个账户的“好友申请接受率”的属性中，因此“好友申请接受率”成为一个极具有信息量和区分度的属性类型，这也直接导致文献 [41] 仅使用几条规则分类器，就可以收到优秀的鉴别效果。然而 CQA 系统中所有的提问者均不能删除虚假用户的回答答案，不能主动阻止虚假账号的强制植入，所以在 CQA 系统中没有像“好友申请接受率”一样具有区分度的属性。同时由于”人工判断“因素的缺失，面向 CQA 系统的虚假账户鉴别准确率一般都不高。

使用分类器来鉴别虚假账户，由第二章可知我们获取了账户主页上的 8 种原始属性，并将直接使用这 8 种原始属性的鉴别结果作为效果的参照基准。基于 CQA 系统给出的原始信息，本文进一步生成了账户 6 个新属性，分别是3.1.1定义的回报率 pro_ans，3.1.2定义的类别熵 $cate\_entropy$，提问类别倾向 $cate\_trend$，经验有效率 ex_ratio，账户问答时间差 $timediff$ 和虚假植入指数 $index\_spammer$。

定义提问类别倾向 $cate\_trend$ 如下：由第三章可知，虚假账户和正常账户活跃的领域差距很大，将虚假账户倾向于提问的类别领域定义为特殊类别集合，类别倾向是指一个账户的所有提问类别中特殊类别所占的比例，类别倾向越大，该用户越有可能属于虚假账户；定义经验有效率 ex_ratio 为：一个用户经验值与提问数与回答数之和的比值。该属性的实际意义为用户的交互行为产生经验的效率。定义账户问答时间差 $timediff$ 为：单个用户提问的所有问题与每个最佳答案张贴的时间差的平均值，即该账户参与的所有问题的问答时间差的均值，以分钟为单位；定义一个账户的虚假植入指数 $index\_spammer$ 为：一个账户所在的一个或多个社团中虚假用户数的比例的平均值，该指数能够表示该用户与虚假账户的亲密性。

在完全回答网络中，9,000 个虚假用户直接连接 42,000 个其他用户，而正常用户直接连接 61,300 个其他类别账户，这里其他类别账户指尚未知道类别的账户，可见虚假账户相比正常账户接触更少的用户，有可能虚假账户内部紧密连接。在章节中，本文指出了虚假账户来源的偏向性，但是由于虚假账户的内聚性，可以通过部分虚假账户粘连出其他模式的虚假账户。在 MSLPA 检测出的 10,745 个账户社团中，至少包含一个虚假账户的社团有 2,307 个，本文称这些社团为被植入社团。图5.1 给出了这 2,000 多个社团中虚假账号的植入概况，

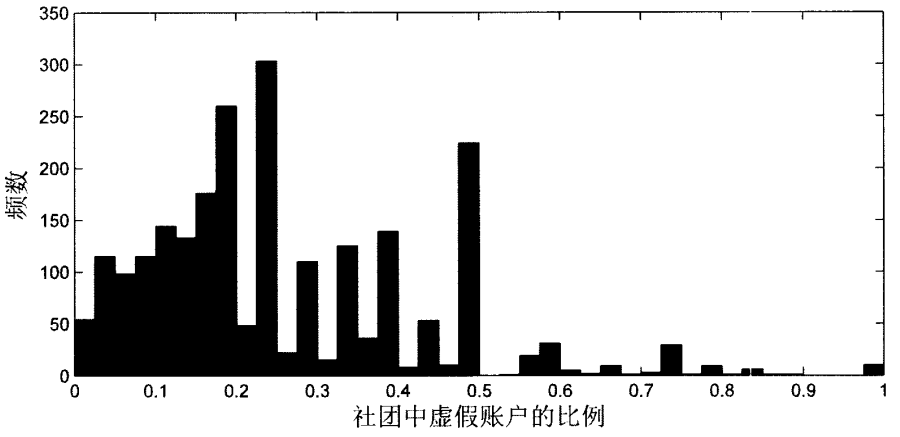

图 5.1 账户社团中虚假账户的比例

横坐标是指每个社团被植入的虚假账户所占的成员比例，纵坐标是指具有对应虚假账户植入比例的社团数，36% 的被植入社团中的虚假账户比例超过 20%，相比而言，初始的 9,000 个正常账户所在的社团中，出现多个正常账户聚集在同一个社团中的可能性较低。

本文使用准确率 *accuracy*[55]，召回率 *recall* [55]，精度 *precision* [55] 和 F_1[55] 度量来衡量不同分类器的鉴别结果。

5.2 鉴别方法的性能分析及结果

我们使用 6,000 个正常账号和 6,000 个虚假账号作为训练集，使用 3,000 个虚假账号和 3,000 个正常账号作为测试集，属性集包含了 7 种原始属性和 6 种生成属性共 13 种属性。weka[56] 是一款免费的，非商业化的，基于 JAVA 环境下开源的机器学习以及数据挖掘软件，本文应用 weka[56] 上的基本分类算法 SVM(SMO)[57]，Logistic[58]，NaiveBayes[59]，AdaBoostM1 [60]，Bagging(J48)[61]，PART[62]，J48[63]，RandomTree[64]，RandomForst[65] 进行分类。

由表5.1可以看出，决策树类的分类器 *accuracy* 要好于其他类型的分类器，其中 J48 分类器算法简单，运行效率高，总体性能好于基于其他决策树的组合分类器。CQA 系统的虚假账户鉴别精度和微博的鉴别精度均处于 78% 与 86%

表 5.1 不同分类器的账号鉴别效果对比

分类器	运行时间	准确率	召回率	精度	F_1
SVM(SMO)	5.7s	78.6%	56.7%	96.2%	71.3%
Logistic	1.65s	80.9%	0.94%	63.4%	75.7%
NaiveBayes	0.08s	79.6%	93.5%	60.7%	73.6%
AdaBoostM1	2s	81.6%	84.4%	74.5%	79.1%
Bagging	5.38s	85.6%	88.6%	79.7%	83.9%
PART	2.16s	84.5%	89.2%	76.1%	82.1%
J48	1.05s	85.4%	85.6%	82.7%	84.1%
RandomTree	1.72s	82.2%	82%	79.5%	80.8%
RandomForst	11.1s	84.3%	85.6%	80.1%	82.8%

之间，但远远小于针对人人网中的虚假账户鉴别精度，这可能是因为前两种社交媒体都没有阻止或者不方便阻止虚假账号植入的功能，例如人人网中的“忽略好友申请”。

本文进一步应用具有优秀鉴别性能的 J48 分类器，来选取账户属性集合中具有较高鉴别能力的属性子集。为精简账户的属性维度，使得鉴别更加高效，本文使用相同的数据，不同的属性组合进行分类对比，一组为原始的“百度知道”爬取到的用户基本属性，一组为基于账户属性和社团属性生成的 6 个新属性，前者称为原始属性数据集，后者称为生成属性数据集。对比原始属性与组合属性的决策树分类效果。对比结果如表5.2所示。

表 5.2 不同的账户属性集合的鉴别结果比较

数据集	叶子数	树规模	准确率	召回率	精度	F_1
原始属性	295	589	75.9%	75.4%	72.4%	73.9%
生成属性	37	73	85.4%	85.6%	82.6%	84.2%

由表5.2可知，生成属性将准确率从 74% 提高到 85%，同时减少了决策树的规模大小。和表5.1相比，账户的属性集从 13 个减少为 6 个，但是分类结果并未降低，同时决策树的规模大大减小，这可能是因为属性之间存在相关性导致信息冗余，产生过拟合现象。本文进一步从决策树提取出几条简易的判断规则如下：

- $index\_spammer <= 0.000884$ AND $cate\_trend <= 0.152672$ AND pro_ans

$<=$ 9.993056 AND $ex\_ratio > 0.001784$: 正常账户

- $pro\_ans > 42$ AND $pro\_ans <= 50.666667$: 正常账户
- $ex\_ratio > 33.666667$ AND $special\_cate <= 0.756303$: 正常账户
- $index\_spammer > 0.027027$: 虚假账户
- $timediff > 759.111111$ AND $pro\_ans <= 24$: 虚假账户
- $pro\_ans > 16.702128$ AND $pro\_ans <= 31$ AND $ex\_ratio <= 23$ AND $timediff > 58.75$: 虚假账户

使用上述 6 条规则对 18,000 个账户实例分类，分类准确率仍然可以达到 80%。在三条识别为正常账户的规则中，可以发现正常账户具有如下共性：虚假植入指数偏小，类别倾向不明显，经验有效率偏高；虚假账户具有虚假指数偏高，账户问答时间差偏大，经验有效率偏低。提问类别熵处于决策树的底层，一般用于对较小的数据集划分，所以上述 6 条规则中并未出现提问类别熵属性。提取出的分类器鉴别规则与第三章中展示的两类用户的统计分析结果一致，两种不同的分析方法所刻画的虚假账户特性相同。

进一步分析账号被预测错误的原因，发现被分类器预测为正常账号的虚假账号具有以下特征类型中的一类：

1. 虚假账号的各项属性值均类似新注册用户，原始属性值均比较小，没有发展为成熟的虚假账号，本文称这类虚假账号为僵尸账号。误判类型的账号集合中，多数属于该种特征；

2. 虚假账号隐秘程度较高，在线查看其在百度知道上的个人信息，发现该虚假账号参与了较多的正常问题的回答，较紧密的植入到百度知道系统中；

3. 标记的虚假账户可能实际上并不是虚假账号，存在样本标记错误，在线查看其在百度知道上的个人信息，没有发现回答记录远远少于回答数的情况，即百度知道官方没有明显得删除该账户的回答信息，且从内容来看也不太像虚假信息。该种误判类型比较少见。

被分类器预测为虚假账号的正常账号具有以下特征中的一类：

1. 正常账号的确是虚假账号，存在样本标记错误。误判类型的账号集合中，约有 7% 属于该种情况。

2. 正常账号的问答信息中存在推销嫌疑，例如有推荐某品牌的回答记录，但是无法确认其推销动机和发帖意图。误判类型账号集合中，多数属于该种特征；

3. 正常账号属于乐于助人的账号，回答了一定数量的虚假问题。

应用到虚假账户鉴别的社团发现结论中，不止虚假植入指数单个属性，实际上还有很多信息没有使用到，例如正常用户暗含在社团中的抵制虚假植入信息，社团间的连接关系等等，在社团层次上仍有很多有价值的属性有待挖掘。

5.3 本章小结

本章将社团发现结果应用于虚假账户鉴别中，给出了用于虚假账户鉴别的属性集合和最佳分类器类型。基于 18,000 个账户实例的实验结果显示，该解决方案可以有效地在大规模 CQA 数据中鉴别出虚假用户。本章进一步分析账户类型误判的原因，这有助于挖掘出更多的虚假账户特征。

第六章　总结与展望

6.1　本文总结

CQA 系统为互联网用户提供优良的问答服务，能够满足用户个性化的需求，及时高效地给出解决方案，CQA 系统也给“专家”用户提供一个共享知识和给予帮助的平台。然而由于 CQA 系统的高度开放性，使得 CQA 系统比较脆弱，容易受到水军，垃圾，广告等不良信息的侵入。CQA 系统有较高的实用性和流行度，因此理解 CQA 系统中用户的行为规律，能够更好更及时地提供 CQA 服务。CQA 系统与其他社交媒体，例如人人网，微博有着本质的差异，因此针对后者的研究成果难以应用到 CQA 系统中。为保证 CQA 系统健康发展，探索 CQA 系统中用户的问答规律，具有重要的应用价值和意义。本文选取百度知道作为 CQA 系统的代表，进行了深入研究，总结研究工作如下：

1. 本文给出一种面向社区问答系统高效率爬虫技术，设计并实现了针对 CQA 系统的网络爬虫程序爬取中文最大 CQA 系统“百度知道”的真实数据。不仅如此，为了后续分析，同时获取了著名虚假交易网站“猪八戒”网站中有关“百度知道”的虚假信息，用以获取虚假样本。针对百度知道设计的网络爬虫可以较容易地修改为其他 CQA 系统。

2. 本文使用初步统计分析方法，探索了问答系统中用户和问题两大主体的属性特征。这些属性具有重要的实际意义。基于真实数据的分析结果，展示了正常账户和虚假账户的属性差异。本文还提出具有一定物理含义的账户属性，例如“回报率”，“类别熵”，“问答时间差”等，这些属性在鉴别虚假账户时具有较高的区分度。

3. 本文给出两种具有不同应用场景的 CQA 网络模型：问答完全网络和回答环形网络，并分析 CQA 网络的属性特征；进一步，本文提出了一个新颖的社团发现算法 MSLPA，该算法能够有效发现有意义的、重叠的、具有层次结构的账户社团，适用于具有大规模数据的 CQA 账户网络。社团发现技术可以用

来为 CQA 用户推荐具有相似兴趣或者相同问题的其他用户，有助于用户可以针对性的问答，提高用户满意度，促进知识的有效传播。在针对中文 CQA 系统的账户行为研究中，本文首次给出并分析了具有不同应用场景的两种 CQA 账户网络模型，并提出一种针对 CQA 系统、简洁高效、可扩展性强的社团发现算法。

4. 本文给出社区问答系统中虚假账户的鉴别方法。作为社团发现技术的一个应用，将社团发现结论应用到虚假账户鉴别中。选取包含 6 个账户属性的具有高区分度的属性集合和简洁高效的分类器，作为 CQA 系统的虚假账户鉴别方案，并验证了该方法的的鉴别有效性。CQA 系统的开放性使得其本身很容易受到虚假信息的植入和污染，而且虚假信息注入量十分庞大，而虚假账户的鉴别方案，有助于阻止虚假信息的扰乱。

6.2 工作展望

本文研究社区问答系统中账户的行为规律，并提出了一种 MSLPA 社团发现算法，将其应用于虚假账户的鉴别，然而研究工作中仍然存在一些问题，具体如下：

1. CQA 账户网络模型的构建中，边的属性有待进一步扩充，权值可以进一步细化分配。MSLPA 算法中标签的传播和节点之间边的权值息息相关，本文中节点间的权值是由节点的问答相遇次数唯一决定，但实际上两个节点之间的亲密度可能和多方面因素相关，例如张贴信息时间差等，所以网络中边的权值设定需要进一步研究。

2. CQA 系统的账户属性需要改善优化，从而继续提高虚假账户鉴别的能力。在虚假账户鉴别的属性提取中，实验结果显示提问类别熵的区分度并不明显，第三章发现虚假账户和正常账户对不同类别的问题关系程度不同，但是提问类别熵不能反映两类用户的关注类别差异，类别倾向属性虽然可以表现关注类别差异，但是也会把具有相似类别爱好的正常用户误认为是虚假账户。因此上述两种属性需要进一步修改。

3. CQA 系统的虚假账户样本来源可以进一步丰富。在虚假交易平台中对虚假账户抽取时，可能会存在一定的偏向性，导致忽略其他模式的虚假账户，

尽管社团发现技术可以弥补部分虚假账户行为模式的缺失，但是若能从数据来源中扩充虚假账户样本，也可以促进在真实 CQA 系统中的虚假账户鉴别。

下一步工作中，我们将重点解决上述问题：优化账户网络模型，从而提高社团发现算法的社团质量；挖掘具有区分度和代表性的账户属性，从而提高虚假账户鉴别的效率。在账户属性挖掘中，为进一步提高虚假账户鉴别的准确率，属性集合可以添加网络属性，例如账户的集聚系数，中介数等，全局属性，例如某属性在某类别中处于的位置信息等。我们预测将用户个体属性，统计属性和网络属性等三种来源的属性结合起来，虚假账号的鉴别准确率会有所提升。

成熟的虚假账号鉴别技术可以用来量化问答页面的可信度，例如通过分析提问者和回答者的“虚假”程度，结合自然语言处理技术，从而得到一个问答页面的真实度和参考价值；同时本文实验结果说明了社团发现技术的引入能够改善在虚假账号鉴别中的效果，这将鼓励我们进一步将社团发现技术服务于更多的应用中，例如新主题检测，推荐账户等。

在目前看来，社区问答系统具有不可替代的功能和作用，必定会继续发展壮大。社区问答系统是一个值得深入研究的对象，探索社区问答系统对全民知识库的扩充和用户行为规律的刻画都具有重要意义。

此页不缺内容

参考文献

[1] Mollá D, Vicedo J L. Question answering in restricted domains: An overview. Computational Linguistics, 2007, 33(1):41–61.

[2] baiduknows. helper. http://baike.baidu.com/view/58034.htm, 2011.

[3] 毛先领, 李晓明. 问答系统研究综述. 计算机科学与探索, 2012, 6(3):193–207.

[4] Shah C, Pomerantz J. Evaluating and predicting answer quality in community QA. Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval. ACM, 2010. 411–418.

[5] Pal A, Konstan J A. Expert identification in community question answering: exploring question selection bias. Proceedings of the 19th ACM international conference on Information and knowledge management. ACM, 2010. 1505–1508.

[6] Yang J, Morris M R, Teevan J, et al. Culture Matters: A Survey Study of Social Q&A Behavior. ICWSM, 2011, 11:409–416.

[7] Logie J, Weinberg J, Harper F M, et al. Asked and Answered: On Qualities and Quantities of Answers in Online Q&A Sites. Proceedings of The Social Mobile Web, 2011.

[8] Pal A, Chang S, Konstan J A. Evolution of Experts in Question Answering Communities. Proceedings of ICWSM, 2012.

[9] Adamic L A, Zhang J, Bakshy E, et al. Knowledge sharing and yahoo answers: everyone knows something. Proceedings of the 17th international conference on World Wide Web. ACM, 2008. 665–674.

[10] 李晨, 巢文涵, 陈小明, et al. 中文社区问答中问题答案质量评价和预测. 计算机科学, 2011, 38(6):230–236.

[11] Pal A, Harper F M, Konstan J A. Exploring question selection bias to identify experts and potential experts in community question answering. ACM Transactions on Information Systems (TOIS), 2012, 30(2):10.

[12] Yang J, Wei X, Ackerman M S, et al. Activity Lifespan: An Analysis of User Survival Patterns in Online Knowledge Sharing Communities. Proceedings of ICWSM, 2010.

[13] Jurczyk P, Agichtein E. Discovering authorities in question answer communities by using link analysis. Proceedings of the sixteenth ACM conference on Conference on information and knowledge management. ACM, 2007. 919–922.

[14] Shtok A, Dror G, Maarek Y, et al. Learning from the past: answering new questions with past answers. Proceedings of the 21st international conference on World Wide Web. ACM, 2012. 759–768.

[15] Jeon J, Croft W B, Lee J H. Finding similar questions in large question and answer archives. Proceedings of the 14th ACM international conference on Information and knowledge management. ACM, 2005. 84–90.

[16] Zhou G, Cai L, Liu K, et al. Exploring the existing category hierarchy to automatically label the newly-arising topics in cQA. Proceedings of the 21st ACM international conference on Information and knowledge management. ACM, 2012. 1647–1651.

[17] Li B, Jin T, Lyu M R, et al. Analyzing and predicting question quality in community question answering services. Proceedings of the 21st international conference companion on World Wide Web. ACM, 2012. 775–782.

[18] 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. ACM, 2012. 783–790.

[19] Agarwal A, Raghavan H, Subbian K, et al. Learning to rank for robust question answering. Proceedings of the 21st ACM international conference on Information and knowledge management. ACM, 2012. 833–842.

[20] Ghosh S, Viswanath B, Kooti F, et al. Understanding and combating link farming in the twitter social network. Proceedings of the 21st international conference on World Wide Web. ACM, 2012. 61–70.

[21] Yang C, Harkreader R, Zhang J, et al. Analyzing spammers' social networks for fun and profit: a case study of cyber criminal ecosystem on twitter. Proceedings of the 21st international conference on World Wide Web. ACM, 2012. 71–80.

[22] Nam K K, Ackerman M S, Adamic L A. Questions in, knowledge in?: a study of naver's question answering community. Proceedings of the SIGCHI conference on human factors in computing systems. ACM, 2009. 779–788.

[23] Anderson A, Huttenlocher D, Kleinberg J, et al. Discovering value from community activity on focused question answering sites: a case study of stack overflow. Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2012. 850–858.

[24] Liu Z, Jansen B J. Factors influencing the response rate in social question and answering behavior. Proceedings of the 2013 conference on Computer supported cooperative work. ACM, 2013. 1263–1274.

[25] Dror G, Pelleg D, Rokhlenko O, et al. Churn prediction in new users of Yahoo! answers. Proceedings of the 21st international conference companion on World Wide Web. ACM, 2012. 829–834.

[26] Welser H T, Gleave E, Fisher D, et al. Visualizing the signatures of social roles in online discussion groups. Journal of social structure, 2007, 8(2):1–32.

[27] Butler B, Kiesler S, Kraut R. Community Effort in Online Groups: Who Does the work and why? Leadership at a distance: Research in technologically-supported work, 2013. 171.

[28] Qi G J, Aggarwal C C, Huang T. Community detection with edge content in social media networks. Proceedings of Data Engineering (ICDE), 2012 IEEE 28th International Conference on. IEEE, 2012. 534–545.

[29] Gargi U, Lu W, Mirrokni V S, et al. Large-Scale Community Detection on YouTube for Topic Discovery and Exploration. Proceedings of ICWSM, 2011.

[30] Lancichinetti A, Fortunato S. Community detection algorithms: a comparative analysis. Physical review E, 2009, 80(5):056117.

[31] Danon L, Diaz-Guilera A, Duch J, et al. Comparing community structure identification. Journal of Statistical Mechanics: Theory and Experiment, 2005, 2005(09):P09008.

[32] Palla G, Derényi I, Farkas I, et al. Uncovering the overlapping community structure of complex networks in nature and society. Nature, 2005, 435(7043):814–818.

[33] Pons P, Latapy M. Computing communities in large networks using random walks. Proceedings of Computer and Information Sciences-ISCIS 2005. Springer, 2005: 284–293.

[34] Donetti L, Munoz M A. Detecting network communities: a new systematic and efficient algorithm. Journal of Statistical Mechanics: Theory and Experiment, 2004, 2004(10):P10012.

[35] Newman M E. Modularity and community structure in networks. Proceedings of the National Academy of Sciences, 2006, 103(23):8577–8582.

[36] Clauset A, Newman M E, Moore C. Finding community structure in very large networks. Physical review E, 2004, 70(6):066111.

[37] Xie J, Kelley S, Szymanski B K. Overlapping community detection in networks: The state-of-the-art and comparative study. ACM Computing Surveys (CSUR), 2013, 45(4):43.

[38] Xie J, Szymanski B K. Towards linear time overlapping community detection in social networks. Proceedings of Advances in Knowledge Discovery and Data Mining. Springer, 2012: 25–36.

[39] Lancichinetti A, Radicchi F, Ramasco J J, et al. Finding statistically significant communities in networks. PloS one, 2011, 6(4):e18961.

[40] techweb. spammer in baidu knows. http://www.techweb.com.cn/prnews/dianshang/archives/4748.html, 2011.

[41] Yang Z, Wilson C, Wang X, et al. Uncovering social network sybils in the wild. Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference. ACM, 2011. 259–268.

[42] Ding Y, Du Y, Hu Y, et al. Broadcast yourself: understanding YouTube uploaders. Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference. ACM, 2011. 361–370.

[43] baiduknows. helper. http://help.baidu.com/question?prod_en=zhidao&class=242&id=1532, 2011.

[44] Koch R. The 80/20 principle: the secret to achieving more with less. Random House LLC, 2011.

[45] Lee Rodgers J, Nicewander W A. Thirteen ways to look at the correlation coefficient. The American Statistician, 1988, 42(1):59–66.

[46] Newman M E. The structure and function of complex networks. SIAM review, 2003, 45(2):167–256.

[47] Newman M E. Assortative mixing in networks. Physical review letters, 2002, 89(20):208701.

[48] Lee S H, Kim P J, Jeong H. Statistical properties of sampled networks. Physical Review E, 2006, 73(1):016102.

[49] cfinder. algrithm of cfinder. http://www.cfinder.org/, 2011.

[50] oslom. algrithm of oslom. http://www.oslom.org/, 2011.

[51] Ahn Y Y, Bagrow J P, Lehmann S. Link communities reveal multiscale complexity in networks. Nature, 2010, 466(7307):761–764.

[52] link. alglink. https://github.com/bagrow/linkcomm, 2011.

[53] Lee K, Caverlee J, Webb S. Uncovering social spammers: social honeypots+ machine learning. Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval. ACM, 2010. 435–442.

[54] Hu X, Tang J, Zhang Y, et al. Social spammer detection in microblogging. Proceedings of the Twenty-Third international joint conference on Artificial Intelligence. AAAI Press, 2013. 2633–2639.

[55] Tan P N, Steinbach M, Kumar V. 数据挖掘导论, 2006.

[56] weka. Data Mining Software in Java. http://www.cs.waikato.ac.nz/ml/weka/, 2013.

[57] Keerthi S S, Shevade S K, Bhattacharyya C, et al. Improvements to Platt's SMO algorithm for SVM classifier design. Neural Computation, 2001, 13(3):637–649.

[58] Meier L, Van De Geer S, Bühlmann P. The group lasso for logistic regression. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2008, 70(1):53–71.

[59] John G H, Langley P. Estimating continuous distributions in Bayesian classifiers. Proceedings of the Eleventh conference on Uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc., 1995. 338–345.

[60] Quinlan J R. Bagging, boosting, and C4. 5. Proceedings of AAAI/IAAI, Vol. 1, 1996. 725–730.

[61] Bauer E, Kohavi R. An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. Machine learning, 1999, 36(1-2):105–139.

[62] Frank E, Witten I H. Generating accurate rule sets without global optimization. 1998..

[63] Chawla N V. C4. 5 and imbalanced data sets: investigating the effect of sampling method, probabilistic estimate, and decision tree structure. Proceedings of the ICML, volume 3, 2003.

[64] Ho T K. The random subspace method for constructing decision forests. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 1998, 20(8):832–844.

[65] Breiman L. Random forests. Machine learning, 2001, 45(1):5–32.

[66] Gogtay N, Giedd J N, Lusk L, et al. Dynamic mapping of human cortical development during childhood through early adulthood. Proceedings of the National Academy of Sciences of the United States of America, 2004, 101(21):8174–8179.

[67] Barrat A, Barthelemy M, Pastor-Satorras R, et al. The architecture of complex weighted networks. Proceedings of the National Academy of Sciences of the United States of America, 2004, 101(11):3747–3752.

[68] Newman M E, Girvan M. Finding and evaluating community structure in networks. Physical review E, 2004, 69(2):026113.

[69] Jiang J, Wilson C, Wang X, et al. Understanding latent interactions in online social networks. ACM Transactions on the Web (TWEB), 2013, 7(4):18.

[70] slpa. algrithm of slpa. https://sites.google.com/site/communitydetectionslpa/ganxis, 2011.

[71] Feng X N, Wu J F, Tian Y. Community Detection based on Social Network Analysis in Question and Answer Systems. Journal of Computational Information Systems, 2013, 9(15):6297– 6305.

致　谢

从 2011 年桂香四溢的秋季来到中科大校园已经近三年了，我有幸成为计算机系学生中的一员，和一群青春洋溢的小伙伴们，接受导师和其他老师的教诲和指导。这三年的时光中，不仅我的专业技能得到迅速提升，我的性格也从毛躁走向稳重，这些收获会成为我人生中重要的财富。借此机会请容许我一一对他们表达真挚的谢意和敬意。

首先要衷心感谢我的导师田野老师。在我学习和研究的过程中，田老师不厌其烦地纠正我作为新生常进入的科研误区，教会我做科研的基本原则和方式方法；鼓励我主动探索感兴趣的领域，给予我自由的发展空间，引导我发现科研道路上的精彩。田野老师知识渊博，治学严谨，教学认真，工作负责，一直在用行动影响着我们，为我们树立学者的榜样。

感谢班主任和计算机系所有的工作人员，他们为我们屏蔽了学习之外的所有纷扰，让我们有机会静心学习和进行科学研究；他们为我们提供方便的研究平台和全面的信息咨询，让我们能够全身心投入到项目工作中去。感谢计算机系的所有授课老师，他们在研究生阶段的指导给我的研究工作打下了基础。

读研的三年是快乐且充实的，这要感谢班级中才华横溢的同学，他们教会我巧妙的学习方法和科研技能，与他们的激烈讨论给予我诸多启迪；还有实验室中可敬可爱的小伙伴们：赵扬，刘邦传，和振华，苏晓东，吴金福等师兄，王东冠，管正等几位同班同学，胥爱芳，石滚，王鑫，谈小冬，何化钧，张欣欣等师弟师妹，是他们给予我无私的帮助和支持，让我顺利度过科研和生活中的各种困难。感谢相伴多年的挚友项泰宁，给予我珍贵的帮助和鼓励。多少年后，我会仍然记得师兄的耐心指点，师妹的乖巧上进，师弟的插科打诨。

感谢中国科学技术大学，这里有一个能激发青年为理想奋斗的环境，有一个为你打开国际视野的平台。这里有兼有暖气和空调的学生宿舍，这里有涵盖八大菜系的食堂，这里有绚烂的樱花和免费发放的枇杷，这里还有深沉的猫咪和安详的汪星人，这就是一个能让人看到世间诸多美好的乐园。

最后感谢我的父母家人，是他们一直在默默地支持者我，给予我强大的后盾。

冯晓楠

2014 年 5 月 23 日

在读期间发表的学术论文与取得的研究成果

研究工作：

1. 国家自然科学基金，基于地理定位的互联网拓扑测量关键技术研究 (61202405)
2. 安徽省自然科学基金，基于视频语义的 P2P 点播：理论与关键技术研究 (11040606Q52)
3. 中央高校基本科研业务费专项基金, 中国大陆大规模 CDN 网络的主动测量研究 (WK0110000024)

已发表论文：

1. Feng X N, Wu J F, Tian Y. Community Detection based on Social Network Analysis in Question and Answer Systems. Journal of Computational Information Systems, 2013, 9(15):6297– 6305. (EI Accession number: 20133516678751)