社会化问答网站使用模式分析
——以雅虎知识堂为例

吴克文 赵宇翔 朱庆华
(南京大学信息管理系 南京 210093)

【摘要】针对新兴出现的社会化问答网站,介绍该类网站的发展和研究现状。以"雅虎知识堂"为例分析其构成要素和运作流程,从计量角度讨论用户使用模式。研究表明,用户在兴趣、参与动机、知识结构和设问方式等方面存在较大差异,总体回答质量不高,为深入理解用户使用行为奠定基础。

【关键词】社会化问答网站 使用模式 长尾理论
【分类号】G203

The Usage Pattern of Social Q&A Site
——Take Chinese Yahoo Answers as an Example

Wu Kewen Zhao Yuxiang Zhu Qinghua
(Department of Information Management, Nanjing University, Nanjing 210093, China)

【Abstract】This paper introduces the development of newly-emerging social Q&A sites and their related studies. By analyzing the elements and operation process of Chinese Yahoo Answers, the usage patterns of social Q&A site are presented. The result shows that there are significant differences in users' interests, usage motivation, knowledge structure and question formulation, while the overall answer quality is not satisfied. This paper lays the foundation for a deeper understanding of usage behavior.

【Keywords】Social Q&A sites Usage pattern Long tail theory

1 引 言

随着互联网的飞速发展,越来越多的用户使用网络搜寻信息以满足自身信息需求。虽然搜索引擎近年来扮演着在线信息服务的重要角色,其他信息获取渠道也不容忽视,例如在线问答服务(Online Q&A Service)。与搜索引擎的信息获取方式不同,在线问答服务允许用户将自己的信息需求以问题形式而不是关键字形式提出,问题解答以小段文字进行展现,而不是以文档列表形式显示。

在线问答服务并不是新近出现的网络服务,服务方式大致可以分为三种:数字参考服务(Digital Reference Service)、专家解答服务(Ask an Expert Service)和社会化问答服务(Social Q&A Service)。虽然这三种服务有共同特点,即都是通过网络进行问题解答,但在服务策略、技术和信息质量控制手段等方面均有不同。数字参考服务和传统图书馆参考服务类似,由图书馆雇用专业搜索人员在线帮助信息用户搜寻有用信息,主要面向有特定任务需求的用户。专家解答服务是迈向社会化的第一步,其运营一般为营利或者非营利性组织,因此在组织性和程序性上弱于数字参考服务。例如,在一些服务系统中,问题的分类直接决定哪一位专家进行解答。社会化问答服务

收稿日期:2009-11-25
收修改稿日期:2009-12-21

应用了 Web2.0 的理念,系统内并没有明确的结构和组织,也没有数字参考服务和专家解答服务所要求的专业人员,它允许用户就任意话题进行发问和解答[1],每个人都会从群体智慧中受益。

社会化问答服务的本质是允许用户以合作的形式分享其他用户的信息,除了简单的问答,用户还可以以评论解答、解答评分、最佳解答投票等方式参与到问题中。在过去数年中,Google Answers, Yahoo Answers, AnswerBag 等一系列社会化问答服务被推向市场,由于其运营成本低、用户门槛低和社会参与性强等特点,迅速获得了大量的用户群,积累了雄厚的社会资本。

2 研究现状

与其他 Web2.0 服务类似,社会化问答服务的核心是用户参与,即用户不仅是信息的搜集或获取方,也是信息的生产、编辑和共享方。社会化问答服务就其研究范畴而言,是用户、信息和技术三者的综合,如图 1 所示:

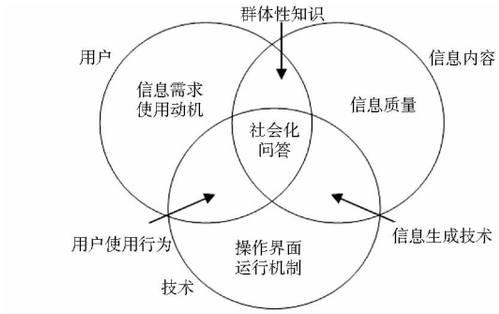

图 1 社会化问答网站研究范畴

国外研究主要分为三种:侧重信息内容、侧重用户和侧重技术:

(1)以信息内容为中心的研究侧重研究解答质量的评价、影响因素、保障机制等。Liu 等人研究发现用户满意度是反映解答质量的重要指标,并提出一系列算法预测用户对解答的满意度,算法涉及问题、解答、用户资料、提问主题和人工评判结果等诸多数据[2]。Haper 等人在预测解答质量的同时,发现社会化问答网站的解答质量竟然高于专业的图书馆参考服务[1]。Kim 等人分析了提问者选择最佳答案的影响因素,认为社会情感类标准(如情感支持、解答者态度、解答者经验、赞同、品味一致和幽默等)具有最强的影响力,其次才是解答效用性和内容价值。

(2)以用户为中心的研究更加多样,包括用户角色、权威用户鉴别和用户信息需求到用户解答时使用的信息源等[3]。Gazan 将解答者分为专家型和综合型,发现提问者喜欢从这两类解答者处同时获得解答,而不是单从一类解答者处受益[4]。Bouguessa 和 Agichtein 等人研究权威用户的认定问题,他们提出了多种用于用户权威等级鉴别的算法[5-7]。

(3)以技术为中心的研究则侧重用户端设计,包括运行机制设计和界面设计等。用户端设计直接影响用户的参与,而用户参与正是社会化问答服务持续进行的关键,Shah 等人通过比较 Yahoo Answers 和 Google Answers 后认为, Google Answers 产品在用户端设计缺少社会化参与机制[8]。

国内目前对社会化问答网站研究较少,主要从信息评价机制和图书馆业务改进两个方面探讨。如余望枝以 BBS 和百度知道为例分析其内在运作方式和相应的信息评价机制[9];赵丽红、毛丹等以百度知道为例分析互动机制、激励模式和服务实效性等方面对于图书馆数字参考咨询工作的启示[10-11]。

本文以雅虎知识堂为例,首先分析社会化问答网站的系统要素与运行流程,从数据计量角度对系统要素进行实证统计分析,发现用户使用行为和信息组织方面的规律,弥补目前国内对于社会化问答网站在计量分析方面的空缺,对进一步研究用户参与动机、系统运行机理、个性化推荐与检索具有重要意义。

3 雅虎知识堂系统要素与运行流程

雅虎知识堂是 Yahoo Answers 的中文版,于 2006 年年初上线运营,是一个任何人都可以对任何问题进行发问和解答的社会化知识共享社区。国内其他社会化问答网站,如搜狗问答、百度知道等站点也都遵从类似的在线交互模型。

如图 2 所示,社会化问答网站系统主要由三大元素组成,分别是:用户、问题和解答。其中,用户分为提问者和解答者两种,另有附着于解答而存在的评论。为更好地进行问题组织,雅虎知识库使用了基于主题的层次分类法,最上层包括计算机网络、娱乐休闲、名人明星、游戏、体育运动、医疗保健、家庭生活、工作理财、交通旅游、投资创业、社会人文、教育学习和科学技

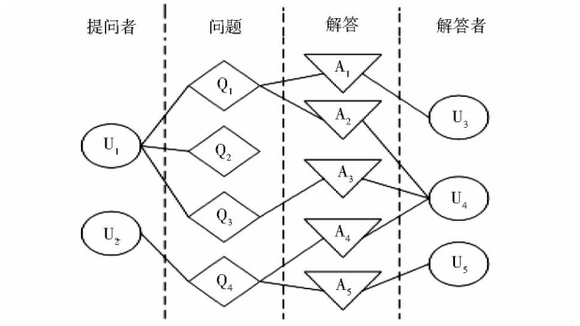

图 2　社会化问答式网站系统模型

术 13 个主题大类,第二层共包括 155 个子类。雅虎知识堂提供积分机制以持续激励用户的参与热情,用户在系统的提问、解答、投票、评价以及知识贡献时均会获得不等的积分,例如提交回答加 2 分,回答被选为最佳答案再加 10 分等;而违反规则的行为和低质量的回答则会扣除部分积分,例如删除回答扣 5 分,已选为最佳答案的回答被投票否定超过 10 次扣 10 分。用户积分的多少除了以分数的形式反映在个人“知识档案”的“知识积分”栏目中,还以“知识等级”的形式从高至低分为圣贤、辅政、仕途、殿试、科举、求学和白丁 7 个等级[12]。积分最高的用户则会出现在“答题王”板块中,积分升降情况可以在“总积分排行榜”中查看。

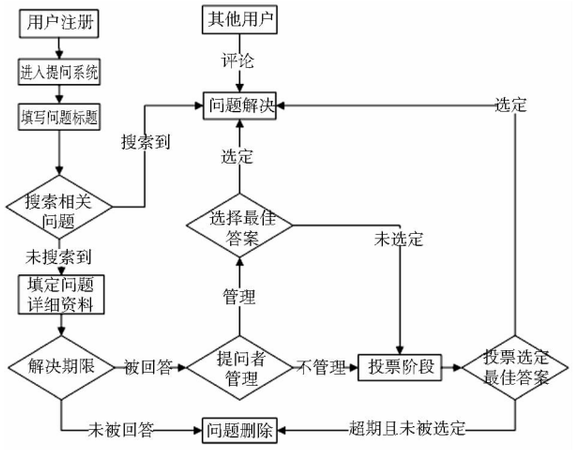

图 3　雅虎知识堂运行流程

雅虎知识堂系统运行流程如图 3 所示,需要提问的用户首先进入提问系统,输入问题标题,如果系统搜索到的相关问题能够解决用户的疑问,用户则停止发问,否则需要输入问题的详细描述,上传相关图片(可选),选择问题分类和悬赏分数,最后设定解答提醒,完成提问流程[13]。一个问题的有效期是 10 天,在这个期限内,其他注册用户可以对该问题进行解答并提供参考资料的链接,如果超过期限且没有人解答,问题将会被系统自动删除[14]。一个问题进入投票流程会有两种情况:当该问题有两个或以上解答并且用户无法选出最佳答案;问题到期时有一个或一个以上解答,但是提问者没有处理问题[15]。当问题交付投票后,会有 5 天的投票期,一般情况下都会在投票期内选出最佳答案。但如果投票期结束时,有两个或两个以上备选答案的票数相同,投票期将会延长 5 天,可延长 3 次。投票期延长 3 次后还未能选出最佳答案,问题将被删除。当最佳答案被选中后,该问题将永久存在于雅虎知识堂知识库中。用户评论有两种途径:对最佳答案提交意见;点击“赞成”或“反对”票,意见和观点票都会附带显示在问题解决历史中[16]。值得注意的是,雅虎知识堂中未解决问题是按时间先后顺序展示在首页的“网友正在问”板块中,意味着在首页上每个问题都有均等的机会被用户注意,而不会因为主题被区分或过滤。

4　数据采集

本文使用的数据采集工具是用 Java 语言自行编制的爬虫程序,循环采集每个主题分类下的页面,使用正则式过滤出问题链接集合,并进一步深入链接分别采集每个问题对应的回答历史列表与关联的用户资料。由于雅虎知识堂并没有与其英文版 Yahoo Answers 一样提供应用程序接口(API),因此对于数据的采集只能通过抓取网站页面进行。雅虎知识堂依据问题状态将问答分为“提问中问题”、“投票中问题”、“已解决问题”、“知识贡献”和“精彩推荐”5 个类别,其中“提问中问题”和“投票中问题”并未完成其从设问到选定最佳回答的全生命周期过程,且“知识贡献”和“精彩推荐”并不涉及问答过程。因此,本文数据采集目标确定为“已解决问题”中的问答条目。“已解决问题”中问答条目的排列是基于时间先后顺序,即新近解决的问题排在首位,每页共 40 个问答条目。值得注意的是,笔者发现在数据试采集阶段,当页数增加到一定数量时,页面数据会报错,因此出于稳定性的考虑,本文的数据集容量确定为按时间倒序排列的前 40% 的问答条目。经过 2009 年 9 月 1 日 - 10 月 25 日近两个月对

“已解决问题”栏目的采集，共收集到问题3 859 970条，占雅虎知识堂总问题数的40%，覆盖该网站近两年以来的全部数据，数据集容量具有统计显著性。数据集共包含与问题对应的回答数17 961 938条，用户数1 848 741个。附带采集的元数据类型包括：问题提出时间、问题解决时间、用户回答时间、用户回答内容长度、用户最佳答案数和用户注册时间等。

同时，本文对“已解决问题”中按主题分类的全部问题数量进行统计，分布如图4所示。截止2009年10月25日雅虎知识堂共有已解决问题9 649 927条，其中问题最多的3个主题分类分别是“家庭生活”、“计算机网络”和“医疗保健”，问题数均超过了100万条，分别占到总问题数的17%，15%和14%。而问题数最少的4个主题分类分别是“交通旅游”、“体育运动”、“明星名人”和“投资创业”，均未超过30万条，分别占到总问题数的3.03%、2.66%、2.07%和0.07%。由此可以看出，雅虎知识堂中的问题在主题分布上存在显著差异。

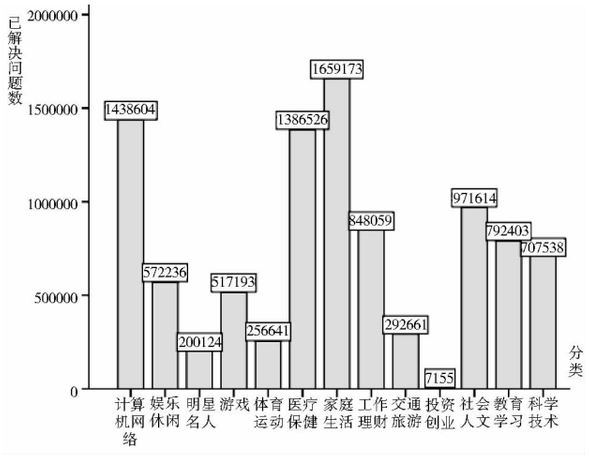

图4　雅虎知识堂“已解决问题”数量分布

5　社会化问答网站使用模式

5.1　用户行为分析

社会化问答系统中用户行为呈现多样化的状态。

（1）从注册用户登录情况进行考察，用户账户年龄定义为从用户注册之日起到本文统计时间点的间隔天数，用户活跃年龄定义为用户注册后提出或回答行为的天数，例如，某位用户在10月1日注册，只在10月3日提出一个问题，则截止10月10日，用户的账户年龄为10天，活跃年龄为1天。使用Pearson相关对用户账户年龄和用户活跃年龄进行相关计算，得出结果为0.379（0.01为显著水平），Pearson相关认为，相关系数0.2－0.4表示正弱相关，即用户账户年龄和用户活跃年龄为正弱相关，说明许多用户注册后并不经常使用该系统，造成活跃年龄相对较低。

（2）考察用户提问数与回答数的相关性，使用Pearson相关计算得出二者相关系数为0.198（0.01为显著水平），表示用户提问数与回答数几乎不相关，说明用户在使用社会化问答网站时带有很强的目的性，有的用户专注于回答问题，而有的用户却只问不答。为了更好地展示用户的提问数与问答数分布，笔者将数据进行自然对数处理并进行散点绘图，结果如图5所示，可以看出用户的提问数与回答数分布呈现明显的长尾现象。大多数用户的提问或回答数量很小，而少部分用户非常活跃，提问或回答集数量相对较大。经过对原始数据进行统计发现，1.57%用户解答了80.9%的问题，而0.29%的用户提出了74.67%的问题。

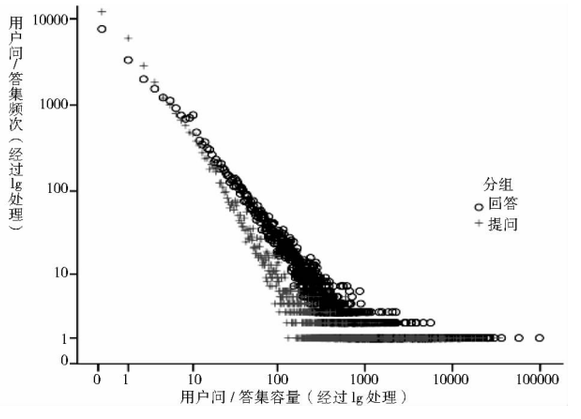

图5　用户的提问/回答集容量分布

（3）为更好地了解社会化问答网站中用户的活跃程度，对问题的生命周期进行统计。根据雅虎知识堂的规则，问题最长的生命周期是30天，但是统计发现，在统计的3 859 970条已解决问题中，极少问题（57个，约占总问题集的0.0015%）会持续到30天才解决。而88.42%的问题都会在提出后5天或者更短时间内解决。如图6所示，通过这两种截然不同的问题解答时间序列曲线可以看出，有的问题在提出后短时间内收到大量回答，迅速被解决，有的问题解答数量的增长则较为平缓。可能原因是：

①提问难度不同，某些问题较为专业，只有少部分用户才能解答，造成被解答几率相对较低，例如询问有关 Java 计算机程序的问题则需要精通该种特定编程语言的用户进行解答；

②问题所处的类别受关注程度不同，或是问题并未出现在页面醒目位置。主题分类的人气程度直接影响接触问题的潜在人数，网站的导航也影响用户对问题的关注，例如出现在首页“网友正在问”板块中的问题会更容易被用户关注。

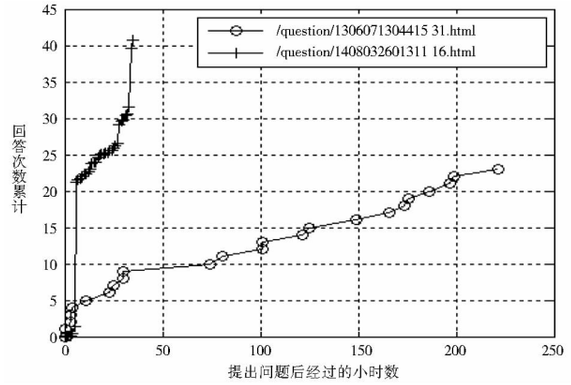

图 6　问题被解答的时间序列

5.2　用户问答分析

图 4 显示了各主题分类的热门程度存在明显不同，深入到各个主题分类的子类设置，考察各主题分类下问题数与子类目数的关系，Pearson 相关显示二者为强相关(0.834，0.01 为显著水平)，即主题子类目数越多，主题分类下问题数也越多。但是也有例外，如“交通旅游”子类目数为 11，而问题数排名倒数第 3。究其原因，“交通旅游”主题的地缘属性较为浓重，例如只有从武汉去过武当山的人可能才能解答哪条路线最优惠。设问者对于收到满意答复的感知期望较低，会阻碍其设问的行为，用户会认为网络上存在更好的主题网站去获取这些信息。至于“投资创业”，其先天性特点也决定了提问数量少，例如具有投资创业思想的人比例不高，投资创业灵感的不稳定性，对在网络上公布投资创业灵感的保密性担忧等。

考察每个分类的回答或问题比，结果如图 7 所示。可以发现“投资创业”、“名人明星”与“家庭生活”的解答或问题比率最高。出现这种现象的原因与分类的自身属性有关，“投资创业”类问题本身就是需要征求群体性意见的问题，“名人明星”则与好奇等人性特点有关，“家庭生活”类是用户共享生活经验和小窍门的场所，多数用户愿意在里面获取知识。回答或问题比最低的是“体育运动”，由于该类问题多涉及确定性知识(如描述某项运动特征)，或者既成事实(如解答某项比赛结果)，答案变种不多，不需要太多次解答即可得到最佳答案。

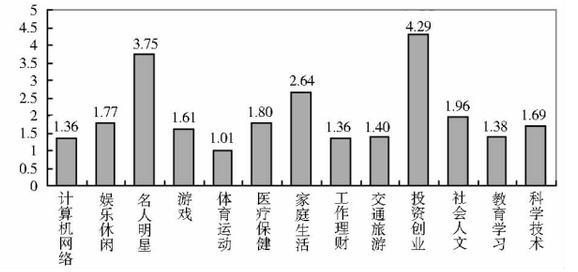

图 7　问题平均解答次数

深入到个人，用户设问方式较为多样，如表 1 所示。社会化问答网站中用户的提问方式与日常生活中提问方式类似，典型的 4 种分别是：

(1)开放式提问，即提出比较概括、广泛、范围较大的问题，对回答的内容限制不严格；

(2)清单式提问，即设问时列出可能的选择项供解答者选择；

(3)假设式提问，即设问时假设一种情况，让回答者在这种情况下作出回答；

(4)判断式提问，即让回答者针对某个问题作是非回答。

设问方式的多样化和口语化会导致社会化问答网站知识库质量不稳定、知识抽取困难、知识获取困难和知识冗余等问题，如果没有很好的方法对目前拥有的庞大知识库进行整理规范，那么该种类型网站演化为更高层级的知识仓库将会变得十分困难。

表 1　雅虎知识堂用户设问方式

设问类型	特点	举例
开放式提问	对解答内容限制不严格	想开个新颖的小店什么好呢
清单式提问	设问式列出可能选项供选择	兼容耗材和原装耗材哪个更适合加墨后使用
假设式提问	回答特定环境下的某种反应	假设生命只剩下最后三天，你会怎么过
判断式提问	针对某个问题作是非回答	CF 卡存储速度会不会越用越慢

从图 5 可以看出用户的回答数量普遍大于提问数量，图 8 则进一步显示了用户提问和回答涉及的类别分布。统计显示，每个用户平均在 2.4 个分类下提问

题,486 647 位用户(约占 26%)只在一个类目下提问题,684 798 位用户(约占 37%)在两个类目下提问题,而只有 213 位用户(约占 0.01%)在全类目下提问题。同样,每个用户平均在 4.4 个分类下回答问题,明显多于提问涉及分类。有 4 000 多个用户在全类目下面解答问题,与提问存在重大反差。用户提问数与提问涉及类目的 Pearson 相关度为 0.411,用户回答数和回答涉及类目的 Pearson 相关度只有 0.320,表明无论用户在提问时还是回答时,都带有较强的随意性,但是用户提问的领域局限性比回答的领域局限性更强。

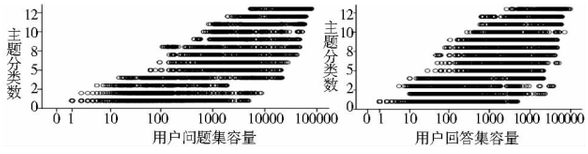

图 8　用户提问和回答兴趣分布

社会化问答网站面临的最大问题就是非领域专家解答的质量及其稳定性问题。对回答字数长度进行粗略统计得出,少于 20 个字的回答占全部回答数量的 18.73%。虽然该统计指标包含正确答案本身较短的情况,但是能从一个侧面反映出回答的整体质量,即社会化问答网站中存在较多无效回答。例如,问题"怎么用手机打长途最省?",回答是"让对方打过来"和"同意楼上的"等信息量不高的语句。再考察用户的积分分布情况,如图 9 所示,可见存在许多用户回答数很高但是积分等级相对较低的情况,说明该用户的回答质量较低,没有获得其他用户认同,也反映出社会化问答网站用户中许多用户并不是某个领域的专家,而是具备多种知识兴趣的普通人,其回答常常过于随意,无法提供高质量的信息。

6　结　语

通过对雅虎知识堂海量数据的分析,本文从计量角度探讨了社会化问答网站的使用模式。研究发现:

(1)用户在使用社会化问答网站时,在使用次数、知识分享方面存在长尾现象明显,说明用户使用该服务的动机和知识结构存在较大差异;

(2)用户的兴趣多样,大部分人设问主要集中在某两个领域,而解答则至少遍布 4 个领域;

(3)用户设问方式多样,直截了当的问题一般解答数较少,开放式问题一般能引起众人讨论;

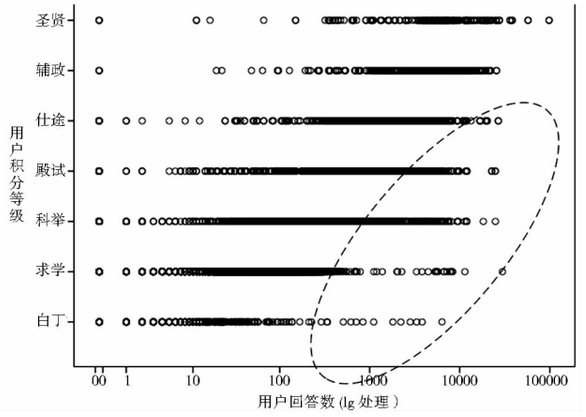

图 9　用户积分等级与回答数分布

(4)总体问答质量不高,无效问答过多。

(5)问题数量的多少主要与该分类自身特性有关。

综上,虽然社会化问答网站虽然具有区别于数字参考服务和专家解答服务的特点,是一种新兴的知识搜寻和获取的工具,但是其需要设计合适的运行机制以支持和改善社会化问答的过程,从满足用户信息需求入手,改进用户寻找已解答问题的功能,改善领域专家解答问题的操作,降低低质量回答(噪声)对系统的干扰。为了达到上述目标,现有的诸多技术可以进行辅助,如采用基于内容挖掘和结构挖掘的个性化机制以帮助用户更好地发现适合的领域专家,采用基于文本分析和链接分析的反垃圾措施以降低系统噪声,以及基于本体或者同义词表的扩展性信息检索帮助用户更有效找到类似的已解决问题等。总之,理解用户使用模式对于科学研究和系统运营都十分必要,本文的研究成果对未来开展更深入的用户和系统机理研究具有重要意义。

参考文献:

[1] Harper F M, Raban D, Rafaeli S, et al. Predictors of Answer Quality in Online Q&A Sites[C]. In: *Proceedings of the 26th Annual SIGCHI Conference on Human Factors in Computing Systems*, Florence, Italy. New York: ACM, 2008:865-874.

[2] Liu Y, Bian J, Agichtein E. Predicting Information Seeker Satisfaction in Community Question Answering[C]. In: *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Singapore, Singapore. New

York: ACM, 2008:483 - 490.
[3]Kim S, Oh S. Users' Relevance Criteria for Evaluating Answers in a Social Q&A Site[J]. *Journal of the American Society for Information Science and Technology*, 2009,60(4):716 - 727.
[4]Gazan R. Specialists and Synthesists in a Question Answering Community[EB/OL]. [2009 - 10 - 25]. http://eprints. rclis. org/8433/1/Gazan_Specialists. pdf.
[5]Bouguessa M, Dumoulin B, Wang S R. Identifying Authoritative Actors in Question - answering Forums: The Case of Yahoo! Answers[C]. In: *Proceedings of the* 14*th ACM SICKDD International Conference on Knowledge Discovery and Data Mining*, Las Vegas, Nevada, USA. New York: ACM, 2008:866 - 874.
[6]Jurczyk P, Agichtein E. HITS on Question Answer Portals: Exploration of Link Analysis for Author Ranking[C]. In: *Proceedings of the* 30*th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Amsterdam, The Netherlands. New York: ACM, 2007:845 - 846.
[7]Jurczyk P, Agichtein E. Discovering Authorities in Question Answer Communities by Using Link Analysis[C]. In: *Proceedings of the* 16*th ACM Conference on Information and Knowledge Management*, Lisbon, Portugal. New York: ACM, 2008:919 - 922.
[8]Shah C, Oh J S, Oh S. Exploring Characteristics and Effects of User Participation in Online Social Q&A Sites[EB/OL]. [2009 - 10 - 29]. http://www. uic. edu/htbin/cgiwrap/bin/ojs/index. php/fm/article/view/2182/2028.
[9]余望枝,朱少强. BBS 论坛与百度知道的信息评价机制探讨[J]. *图书馆学研究*, 2008(12):81 - 83.
[10]赵丽红. 互动式知识问答分享平台对虚拟参考咨询服务的启示[J]. *图书馆建设*, 2009(5):62 - 64.
[11]毛丹. 中文网络知识问答平台对数字参考咨询服务的启示[J]. *图书馆学研究*, 2009(6):79 - 81.
[12]雅虎知识堂帮助中心. 知识等级表[EB/OL]. [2009 - 10 - 29]. http://ks. cn. yahoo. com/info/ranking. html.
[13]雅虎知识堂帮助中心. 如何提交问题[EB/OL]. [2009 - 10 - 29]. http://help. cn. yahoo. com/answerpage_1601. html.
[14]雅虎知识堂帮助中心. 如何回答问题[EB/OL]. [2009 - 10 - 29]. http://help. cn. yahoo. com/answerpage_1616. html.
[15]雅虎知识堂帮助中心. 如何将问题设置为投票[EB/OL]. [2009 - 10 - 29]. http://help. cn. yahoo. com/answerpage_1623. html.
[16]雅虎知识堂帮助中心. 如何进行评价[EB/OL]. [2009 - 10 - 29]. http://help. cn. yahoo. com/answerpage_1642. html.

（作者 E - mail:kewen - wu@163. com）