社区问答系统研究综述

张中峰 李秋丹
(中科院自动化研究所 北京 100190)

摘 要 作为一种新兴的知识共享模式,社区问答系统(CQA)具有交互性、开放性的特点,能够更好地满足为用户提供个性化的信息服务的需求。对社区问答系统的研究及应用现状进行综述,系统阐述了用户行为模式、内容质量检测、问题检索等 CQA 中主要问题的研究以及 CQA 在其他媒体中的应用。最后展望了 CQA 中下一步值得研究的问题。本讨论有助于进一步丰富和拓展 CQA 的研究。

关键词 社区问答系统,用户行为模式,内容质量检测,问题检索
中图法分类号 TP39,TP31 **文献标识码** A

Studies on Community Question Answering—A Survey

ZHANG Zhong-feng LI Qiu-dan
(Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China)

Abstract As a burgeoning platform for knowledge sharing, community question answering (CQA) is capable of satisfying personalized information needs for individuals, with its distinct type of interactions among users and openness. This paper reviewed the researches and applications of CQA. We systemically described the research topics in CQA, such as user behavior analysis, content quality detection, question searching etc, and the application of CQA in other forms of social media. Finally, we discussed some issues for further study. The discussions in this paper are beneficial to enrich and expand researches in CQA.

Keywords Community question answering, User behavior, Content quality detection, Question search

以用户为中心的 Web2.0 更加注重系统中用户的交互作用,用户既是网络内容的消费者,又是内容的提供者。目前,比较流行的 Web2.0 应用包括 Flickr 和 YouTube 等多媒体共享网站、Facebook 等社交网站、Delicious 等社会标注系统。以 Yahoo!Answers 和百度知道等为代表的社区问答系统通过提供问题和答案的形式将有信息需求的提问者和乐于分享知识的回答者直接关联在一起,并允许用户对已有内容进行评价和检索,从而提供了相对搜索引擎更加直接和有效的信息获取方式。

1 引言

社区问答系统(Community Question Answering, CQA)以其灵活的用户交互特性,能够满足人们获取和分享知识的需求,而逐渐成为广受用户喜爱的知识共享平台。从 2006 年到 2008 年,全球 CQA 网站的用户流量增长了 800%[1]。到 2009 年 12 月,中文 CQA 平台百度知道已收藏了 7 千多万已解决的问题(http://zhidao.baidu.com/)。腾讯旗下的搜搜问问则已收录上亿条已解决问题,同时在线人数超过 200 万(http://wenwen.soso.com/)。CQA 为搜索引擎的发展提供了丰富的内容资源,促使其提供更加个性化的搜索服务。统计显示,2008 年 2—7 月百度搜索的月访问量有超过 12%来源于百度知道,而搜搜问问对腾讯搜索的月访问次数贡献率 2008 年 7 月更是达到 30.1%[2]。

与其他社会媒体相比,CQA 提供了一种特有的交互方式。首先,提问者将其信息需求以问题的方式提交给系统,并等待其他用户给出答案。回答者根据其个人兴趣、知识水平,选取适当的未解决问题来回答,以分享自己的知识。然后提问者对得到的答案给出评价并选取最满意的一个作为最佳答案。如果提问者无法对答案的质量给出评价,可以将问题送入投票系统,由其他用户投票选出最佳答案。CQA 系统将已解决的问题及其答案存档并提供搜索功能,供外部用户检索。因此,用户在 CQA 中可以扮演 4 种不同的角色:提问者、回答者、评价者和搜索者。CQA 由用户、问题和答案 3 种基本元素组成。图 1 描述了 CQA 中用户与系统的交互过程。

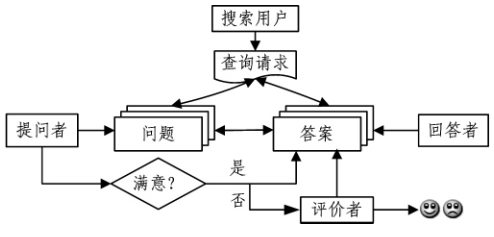

图 1 CQA 用户交互过程示意

CQA 的成功及其独特的交互方式,引起了研究者们极大

到稿日期:2009-12-17 返修日期:2010-03-16 本文受 973 国家基金项目(2007CB311007),国家自然科学基金(60703085)资助。
张中峰(1984—),男,博士生,主要研究方向为信息检索等,E-mail:zhongfeng.zhang@ia.ac.cn;李秋丹(1976—),女,副研究员,主要研究方向为信息检索等。

的研究兴趣。本文系统阐述了 CQA 的主要研究问题和研究现状，在深入分析用户模式的基础上，系统阐述了内容质量检测、问题检索等研究问题，并讨论了 CQA 在其他媒体中的应用。最后对 CQA 未来的研究问题进行了展望。

2 社区问答系统的相关研究

作为一种新兴的知识共享模式，社区问答系统已经吸引了大量用户参与，并存储了大量已解决的问题，以满足用户进一步的信息需求。然而，日益增长的用户和数据资源也为 CQA 研究提出了许多挑战。理解用户行为模式，准确定位用户需求，为用户的查询请求提供高质量的答案等成为 CQA 研究中亟待解决的问题。

2.1 用户行为分析

用户行为分析主要研究 CQA 社区中用户、问题和答案 3 个基本元素之间的统计关系，包括用户活跃性分布规律、用户动机、CQA 问题所反映的用户需求特点等，从而为用户提供更好的信息服务提供依据。文献[3]将 CQA 中的问题分为事实类、观点类或两者综合 3 类，并分析了 CQA 中这 3 类问题随时间的变换规律，表明 CQA 更有利于解决观点类和综合性问题，而对于事实类问题随问题数的增加，答案质量逐渐降低。非事实类问题通常能得到更多答案，用户活跃程度更高[5]。

对 Yahoo！Answers10 个月的数据分析表明，CQA 中用户提交的问题数、答案数和最佳答案数以及每个问题得到的答案数均服从幂律分布。70％的注册用户会提交问题，50％的用户会回答其他用户提交的问题，而只有 23％的用户会对已有的内容进行投票。所有已解决问题的最佳答案中，有 47％为提问者选出，31％的最佳答案得到的投票包含回答者本人的一票，说明系统中很大一部分最佳答案的质量是值得商榷的。用户通过问答关系形成的用户网络为强连通图[4]。

文献[6]将用户参与 CQA 的问答过程的动机分为利他主义、学习、商业目的、自我能力证明和业余爱好等，而用户希望通过 CQA 获得如常识性知识、时事、领域知识、建议和观点等知识。文献[8]分析了 CQA 中的问题所反映出的不同类型的用户意图(社会化意图、个人观点等)，提出了一系列特征来识别具有社会化意图的问题。对百度知道评价机制的分析表明，CQA 社区中存在自动发生的评价机制[9]。

2.2 内容质量检测和专家用户发现

CQA 的开放性使一个用户可以直接从其他用户那里获取所需要的信息，从而更准确地满足其信息需求。然而，由于缺乏控制措施，由用户编辑的内容在质量上的差异较大。对 CQA 中答案质量的分析表明，大约 13％的答案为无价值的答案，更有大约 1％的答案存在作弊行为[7]，同时内容质量与用户的活跃程度和权威度存在密切关系。因此，从用户编辑的内容中检测出高质量的内容及对用户的权威度评价的方法研究是 CQA 成功的基础。

目前，内容质量检测通常被描述为分类问题，所抽取的特征包括内容的文本特征、用户关系特征和 usage 特征等非文本特征。常用分类器包括决策树[11]、最大熵模型[12]以及层次分类器[13]。上述分类方法将问题和答案看作相互独立的部分，分别抽取特征及进行质量检测，忽略了问题和答案之间的紧密联系。文献[14]提出一种类比推理方法衡量问题和答案的相似度，并用于检测高质量的答案。

专家用户发现是根据权威度对用户进行排序。文献[16]将“专家”定义为回答过相似问题的用户，并将专家发现看作传统的信息检索(IR)问题，分别研究了查询相似度(QL)模型、关联模型(RM)和基于聚类的语言模型(LM)在专家发现中的性能。文献[15]根据用户之间的问答关系构建用户之间的网络结构，并运用 HITS 算法对用户网络进行链接分析，计算用户权威度。文献[10]将 CQA 中用户、问题及答案之间的关系表示为二分图结构，并将图模型与回归模型相结合，提出了一种半监督学习方法 CQA-MR，同时检测内容质量和用户权威度。

CQA 系统通常依赖用户对问题或答案的投票等参与方式确保内容的质量。然而有超过 31％的用户会将自己的答案选为最佳答案[4]，从而降低了用户反馈的质量。文献[17]将用户恶意进行投票的行为称为“投票作弊”(vote spam)，并区分了加性投票作弊和减性投票作弊。文中假设投票作弊行为服从高斯分布，从而提出了 GBrank-robust 算法，在答案检索中检测作弊行为。

2.3 问题检索、推荐及主客观性判断

本小节针对 CQA 中的问题进行研究和总结，包括相似问题检索、问题推荐以及问题的主客观性判断。

相似问题检索是指针对用户的查询请求，从 CQA 历史记录里检索与之相似的已解决问题，并将这些问题的答案推荐给用户，从而避免了用户重复提交问题，也方便用户更加快速地获取问题的答案。

Song 等将问题的效用(utility)定义为该问题被其他用户重复提问的概率，认为问题的有效性取决于问题本身，而与查询词无关，从而将问题搜索看作一种静态排序[19]。文章将 n 元语言模型与 LexRank 算法相结合，对问题按有效性进行排序。然而，这一做法并没有考虑用户的查询词所反映出的需求信息。

Duan 等认为一个问题由主题(topic)和主题焦点(topic focus)两部分组成，相似问题检索是查找与用户查询问题具有相同主题焦点的问题。并用最小描述长度(MDL)树识别问题的主题和主题焦点，根据问题的主题焦点进行相似问题查找[18]。

Jeon 等基于 IBM model1 提出了一种翻译模型[20]，用于计算两个问题之间的语义相似度，从而进行相似问题检索。文献[21]进一步考虑了检索到的问题所得到答案的质量，为用户的查询请求提供相关并具有高质量答案的相似问题。Jeon 等人的翻译模型没有考虑噪声控制，在模型建立过程中容易将一些并不重要的词，如停滞词等引入模型，从而降低了检索性能。为克服这一问题，Lee 等提出在构建翻译模型前，首先对所有词按权重进行排序，并滤除权重较低的词[22]。与统计翻译模型不同，文献[24]提出从现有的问题和答案集合中抽取词的共现信息，并根据共现信息计算不同词的相关性，进而提高相关问题检索质量。

文献[23]用语义树结构进行相似问题匹配，并用实验证明语义树比 bag-of-word 和传统的树方法更适合相似问题检索。

以上研究基本是仅考虑查询内容的语义信息，较少利用 CQA 系统本身提供的信息。文献[26]利用 CQA 系统的类别

结构进行相关问题搜索。文献[25]在问题检索中将用户的投票信息考虑在内，并针对用户投票中存在的作弊等噪声问题，提出了一种“majority-based perceptron”算法。

当用户提交一个查询问题时，对问题的主题非常清楚，但经常无法准确描述问题是关于该主题的哪些方面，或者并不清楚该主题在多个方面与用户的信息需求相关。Cao 等将主题的不同方面(aspect)称为主题焦点(topic focus)，在用 MDL 树区分问题的主题和主题焦点的基础上，研究向用户的查询问题推荐属于同一主题但不同主题焦点的问题，帮助用户更全面地把握问题[27]。

尽管 CQA 吸引了大量用户参与，但仍有许多问题没有得到回答。在已解决的问题中，很大一部分答案的质量并不能让用户满意。只有 20％的用户回答的问题数超过 5 个[35]。造成这一现象的原因之一在于许多用户，即便是有经验的用户，也很难从大量新提交的问题中找到感兴趣并有能力解决的问题。在 CQA 中，用户通常需要浏览问题的分类结构，从成千上万的开放问题中找到感兴趣的问题，这一耗时费力的过程打消了许多用户参与的积极性。为了解决这一问题，问题推荐自动将恰当的问题推荐给合适的用户，节省回答者找到新问题的时间，同时帮助提问者更快获得高质量的答案。

Guo 等人将主题特征和基于词的特征相结合，提出了一种产生式概率模型用于问题推荐[35]：首先，用 AT-LDA 模型抽取社区主题，并将问题、答案及用户属性表示为主题的分布；其次，将新提交的问题表示为主题特征和基于词的特征，并推荐给合适的用户。当用户希望回答问题以分享自己的知识时，文献[36]研究了用 pLSA 模型从未解决问题集合中选取用户可能感兴趣的问题并推荐给用户。基于用户先前的提问和回答的行为，文献[38]提出了 3 种不同的语言模型方法来表示用户的权威度，并将新提交的问题推荐给对问题主题感兴趣的、权威度比较高的用户。基于问答过程中形成的用户网络结构，文献[31]研究了基于加权 HITS 算法的问题推荐方法。

上述方法没有考虑用户的工作量和问题的优先级。事实上，用户一次所能解决的问题数是有限的，而为了均衡提问者的信息需求，较早提交的问题需要有更高的优先级，以减少提问者等待的时间。文献[37]提出问题推荐需要综合考虑用户的兴趣和权威度、向用户推荐问题的数目以及问题的优先级。

不同类型问题所得到答案的质量、用户参与的活跃程度等也表现出不同的特点[3,5]。答案摘要生成的研究表明，不同类型问题的答案具有不同的特点，需要采用不同的摘要生成技术[39]。了解用户提交问题的意图，针对不同类型的问题开发不同的检索算法，也有助于进一步提高问题检索、答案检索等的检索精度。Li 等人区分了主观性问题和客观性问题[40,41]。客观性问题通常期望得到的答案中包含可靠的权威信息，而主观性问题所要求的信息往往包含个人观点、判断和经验等。

Li 等人将主客观性判断描述为分类任务，并对每个问题及其相应的答案提取了 TF 特征、n 元语法特征和 POS 特征等文本特征。文献[41]采用 SVM 对问题的主客观性进行分类，并比较了不同类型的特征对分类效果的影响。然而 SVM 等分类算法需要大量的标注数据，而人工对数据进行标注是一项耗时费力的工作。为了使用较少的标注数据训练分类器，文献[40]利用了未标注样本中所包含的信息，采用半监督学习算法 Co-training 进行主观性判断。

2.4 答案检索、摘要生成及用户对答案满意度的预测

除了答案质量检测，针对答案进行的研究还包括答案检索、答案摘要生成及用户对答案满意度的预测等。我们在本节分别讨论这些问题。

同样是针对用户提交的新的查询问题，从历史数据中检索相关内容，以快速满足用户的信息需求，与相似问题检索的区别在于，答案检索直接向用户推荐满足其查询需求的高质量的答案。相似问题检索的研究中，较少考虑答案本身的内容和质量，而答案检索通常同时考虑问题和答案。

文献[28]使用 GBRank 排序算法，对每个查询问题、问题-答案(Q-A)对中的问题和答案抽取了一系列特征，包括文本特征、统计特征(如长度)等，以及相关用户的特征，如用户提交的问题数、答案数和提供的最佳答案数，进而对相关的 Q-A 对进行排序。文献[17]进一步提出 GBRank-robust 算法，以消除投票作弊对排序结果的影响。Tu 等[30]在对 Q-A 对的答案进行排序过程中，引入了模拟退火算法，以取得更好的排序效果。

文献[29]从 CQA 历史数据中得到相关 Q-A 对后，进一步考虑了这些答案的质量及其与用户查询问题的相关性，将与用户查询请求相关的高质量的答案推荐给用户。在评估答案质量时，作者同时考虑了答案的文本特征和回答者的权威度。

在答案检索过程中，通常认为在答案集合中与一个答案相似的回答数越多，该答案越容易被用户接收，从而在排序中权重越高。Achananuparp 等人认为，这类答案恰恰因为重复出现而成为冗余内容，并不能向用户提供更多信息。因此，在排序过程中引入了惩罚函数来降低这类冗余答案的权重[32]，从而使非事实类问题可以得到更加多样化的回答。

文献[31]研究了搜索到的 Q-A 对的聚类算法，以改善搜索结果的表现形式，帮助用户快速定位信息。

现有 CQA 服务中，通常一个问题有唯一的最佳答案，并且该最佳答案由提问者从答案集合中选出或者由社区用户投票产生。Liu 等人的分析表明，尽管多数选出的最佳答案是可重用的，其中 48％并不是唯一的最佳答案[39]。为了充分利用其他用户给出的回答，更全面满足提问者的信息需求，文献[39]提出对问题的所有答案进行摘要生成来代替单一的最佳答案，并针对观点性的问题和开放性的问题提出了不同的摘要生成方法。实验表明，对这两类问题，摘要生成技术所产生的答案比单一的最佳答案更能满足用户的信息需求。但可读性仍然是摘要生成技术所面临的难题之一。

在 CQA 中，如果问题的提问者对其他用户给出的答案感到满意，通常会将这些答案选为最佳答案或给出其他反馈信息。而满足众多用户的信息需求，正是 CQA 能够吸引更多用户参与，保持其持续增长的关键所在。因此，给定一个问题及其相应的答案集合，预测提问者对社区其他用户给出的答案的满意度，成为 CQA 研究的另一个主要任务，有助于进一步了解用户的信息需求，把握用户意图，并对答案进行排序。

Agichtein 等人认为，如果提问者自己选出了最佳答案并关闭问题，或者对答案给出了三星以上的评价，则该提问者的

信息需求得到了满足[33]。文献[33,34]提出用分类算法来预测用户的满意度,比较了决策树、SVM、Adaboost 和朴素贝叶斯等多种分类器在预测用户满意度问题中的性能。每个样本的特征向量由 1000 个文本特征和 72 个非文本特征组成,其中非文本特征包括问题相关统计特征、问题与答案的关系特征、提问者相关特征、回答者相关特征以及类别特征等。

3 社区问答系统在其他媒体中的应用

论坛作为一种在线交流平台,比 CQA 有着更悠久的历史和丰富的用户产生的数据。对 40 个论坛的分析表明,90% 的论坛包含问答知识[42]。因此从论坛中抽取问答知识有助于丰富 CQA 的数据资源,从而吸引更多的用户参与。与 CQA 不同,论坛中一个合集(thread)可能包含多个问题及这些问题的答案。因此,论坛中问答关系的抽取通常包含问题识别和答案检测两个子任务。问题识别的任务是检测出一个合集中所有的问题,通常被看作分类问题。文献[42]从每个帖子的文本中抽取序列模式作为分类器特征,并用 Ripper 分类算法检测问题;文献[49]比较了不同类型特征对问题识别结果的影响,证明非文本特征对提高检测性能起到关键作用。答案检测的任务是对检测出的每个问题,从同一合集中找出相应的答案信息。文献[49]采用了与问题识别同样的分类方法检测答案。而文献[42]采用了一种基于图论的无监督排序算法,将每个候补答案看作图中的节点,两个答案的相似度由其文本相似度、答案与问题的相似度及答案对应用户的权威度 3 种因素共同决定。

数字参考咨询(DRS)是图书馆情报机构的核心服务之一。社区问答系统的出现,为 DRS 的发展在理念、技术和应用等方面提供了借鉴意义[44,45]。而 DRS 与 CQA 的合作有利于双方的共同发展[43]。

现有的 CQA 系统是基于文本的,问题和答案全部用文字形式描述。随着 Web2.0 的发展,多媒体信息逐渐成为流行的信息交流方式。“一图胜千言”,对许多问题,答案如果以图像或视频等可视化的方式给出,则更直观形象并易于理解,也更能提高用户参与的积极性。基于问答系统和图像匹配技术,Yeh 等提出了一种基于图像的问答系统[46]。该系统由 3 层结构组成:第一层为模版匹配层,允许用户提交图像查询请求并从网上检索相匹配的图片;第二层为信息检索层,存储已解决的基于图像的问题并允许用户从中检索相关答案;第三层为人机交互层,允许用户提交基于图像的问题和答案。Li 等人研究了基于 YouTube 的视频问答系统,根据用户的查询请求从 YouTube 丰富的视频资源中检索相关的视频作为答案[48]。视频问答系统由两步组成,第一步进行相似问题查找,扩展查询问题的覆盖率,使问题得到更全面和准确的描述;第二步为视频排序,从 YouTube 中检索相关视频,并采取视频分析、观点分析和视频去重等策略,对相关视频进行排序。随着移动设备的逐渐普及,越来越多的网络应用开发了移动终端,以随时随地满足用户的信息需求。Yeh 等研究了基于图像的问答系统在移动设备中的应用[47]。

4 小结与展望

随着网络技术的发展,社区问答系统已成为一种新兴的知识共享平台。用户可以根据自己的信息需求提出问题,并由其他用户给出解答,相对搜索引擎等具有更好的用户交互特性,得到了众多用户的青睐。本文对 CQA 的研究现状进行了总结性介绍。尽管国内外对 CQA 进行了研究并取得了很好的成果,然而作为一个新兴的研究主题,CQA 仍存在一些尚未解决的问题需要进一步的研究:

1)个性化及社区化是 Web2.0 技术的主要特点,为不同背景、不同兴趣的用户提供个性化服务得到越来越广泛的研究。在 CQA 中,如何为用户提供个性化的搜索和推荐服务,以满足其特定的信息需求是 CQA 进一步研究的方向之一。

2)越来越多的 Web2.0 应用提供了丰富的在线资源。尽管从论坛中抽取问答知识已经得到初步研究,CQA 在其他 Web2.0 中的应用仍未受到重视。如 twitter 等微博客的兴起促进了即时搜索的发展,极大缩短了人们获取新知识的时间。CQA 与即时搜索结合,可以实时性地满足用户的信息需求。

3)CQA 用户编辑内容的形式为人们获取和分享信息提供了便利,但同时为一些用户不合理地使用 CQA 资源提供了途径。尽管 Yahoo!Answers 与百度知道等允许用户对社区中出现的不良信息及广告、刷分等作弊行为进行投诉,以此保证社区的健康发展,但开发适当的策略以自动滤除这些恶意信息,无疑会帮助 CQA 提供更好的用户体验。

4)目前 CQA 中的研究主要从单个用户角度出发,以满足其获取或分享知识的需求。研究 CQA 用户交互网络中的社区结构及其演变规律以及社区中用户感兴趣主题的演化规律,有助于进一步了解 CQA 的特点并开发更多样化的应用。

5)CQA 丰富的用户资源为开发商业应用提供了可能,人们已开发出了许多成功的商业应用案例[50,51]。但 CQA 的商业价值及其可能的商业模式、CQA 对用户商业行为的影响等,在学术界的研究仍较少。

6)CQA 当前的研究多是基于 Yahoo!Answer 等英文问答系统。而百度知道等中文问答系统也取得很大成功并可能表现出不同的特点,因此中文社区问答系统的应用也是未来的研究方向之一。

结束语 社区问答系统已经成为一种重要的知识共享平台。随着网络资源的日益丰富和用户信息需求的多样化,向用户提供及时准确的信息,是社区问答系统发展的基础,也是相关研究的出发点。本文系统阐述了 CQA 的研究现状,并对 CQA 研究中存在的问题和将来的研究趋势进行了探讨。本文的讨论有助于进一步丰富和拓展 CQA 的研究。

参 考 文 献

[1] Biz Report. Hitwise: Question/Answer website business is booming[EB/OL]. http://www.bizreport.com/2008/03/hitwise_questionanswer_website_business_is_booming.html#comments, March 21, 2008

[2] http://news.iresearch.cn/viewpoints/84557.shtml

[3] Liu Y, Agichtein E. On the Evolution of the Yahoo! Answers QA Community[C]// the ACM SIGIR International Conference on Research and Development in Information Retrieval. Singapore, 2008: 737-738

[4] Gyöngyi Z, Koutrika G, Pedersen J, et al. Questioning Yahoo! Answers[C]// First Workshop on Question Answering on the Web at the 17th International World Wide Web Conference. Beijing, China, 2008

[5] Lada A A, Zhang J, Bakshy E, et al. Knowledge Sharing and Yahoo Answers: Everyone Knows Something[C]// Proceeding of the 17th International Conference on World Wide Web. Beijing, China, 2008: 665-674

[6] Nam K K, Ackerman M S, Adamic L A. Questions in, Knowledge iN? A Study of Naver's Question Answering Community [C]// Proc. of CHI'09. Boston, MA, 2009: 779-788

[7] Su Q, Pavlov D, Chow J, et al. Internet-scale collection of human-reviewed data[C]// Proc. of WWW2007. Banff, Alberta, Canada, 2007: 231-240

[8] Rodrigues E M, Frayling N M. Socializing or knowledge sharing?: characterizing social intent in community question answering[C]// Proc. of CIKM 2009. Hong Kong, China, 2009: 1127-1136

[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: 81-87

[10] Bian J, Liu Y, Zhou D, et al. Learning to Recognize Reliable Users and Content in Social Media with Coupled Mutual Reinforcement[C]// Proceedings of the 18th International World Wide Web Conference. Madrid, Spain, 2009: 51-60

[11] Agichtein E, Castillo C, Donato D, et al. Finding High Quality Content in Social Media[C]// Proc. of the International Conference on Web Search and Web Data Mining. Palo Alto, California, USA, 2008: 183-194

[12] Jeon J, Croft W B, Lee J H, et al. A Framework to Predict the Quality of Answers with Non-textual Features[C]// Proceedings of the 29th International ACM SIGIR Conference. Seattle, Washington, USA, 2006: 228-235

[13] Blooma M J, Chua A Y K, Goh D H L. A predictive framework for retrieving the best answer[C]// Proceedings of SAC2008. Fortaleza, Ceara, Brazil, 2008: 1107-1111

[14] Wang X, Tu X, Feng D, et al. Ranking community answers by modeling question-answer relationships via analogical reasoning [C]// Proc. of SIGIR 2009. Boston, MA, USA, 2009: 179-186

[15] Jurczyk P, Agichtein E. Discovering Authorities in Question Answer Communities Using Link Analysis[C]// ACM Conference on Information and Knowledge Management (CIKM). Lisbon, Portugal, 2007: 919-922

[16] Liu X, Croft W B, Koll M B. Finding experts in community-based question-answering services[C]// Proc. of CIKM 2005. Bremen, Germany, 2005: 315-316

[17] Bian J, Liu Y, Agichtein E, et al. A Few Bad Votes Too Many? Towards Robust Ranking in Social Media[C]// The 4th International Workshop on Adversarial Information Retrieval on the Web. Beijing, China, 2008: 53-60

[18] Duan H, Cao Y, Lin C Y, et al. Searching Questions by Identifying Question Topic and Question Focus[C]// Proc. of ACL: HLT 2008. Columbus, OH, 2008: 156-164

[19] Song Y, Lin C Y, Cao Y, et al. Questing Utility: A Novel Static Ranking of Question Search[C]// Proc. of AAAI 2008. Chicago, Illinois, USA, 2008: 1231-1236

[20] Jeon J, Croft W B, Lee J H. Finding Similar Questions in Large Question and Answer Archives[C]// Proceedings of the ACM Fourteenth Conference on Information and Knowledge Management. Bremen, Germany, 2005: 84-90

[21] Xue X, Jeon J, Croft W B. Retrieval models for question and answer archives[C]// Proc. of SIGIR 2008. Singapore, 2008: 475-482

[22] Lee J T, Kim S B, Song Y I, et al. Bridging Lexical Gaps Between Queries and Questions on Large Online Q&A Collections with Compact Translation Models[C]// Proc. of EMNLP 2008. Honolulu, Hawaii, USA, 2008: 410-418

[23] Wang K, Ming Z, Chua T S. A syntactic tree matching approach to finding similar questions in community-based QA services[C]// Proc. of SIGIR 2009. Boston, MA, USA, 2009: 187-194

[24] Lee J T, Song Y I, Rim H C. Computing Word Semantic Relatedness for Question Retrieval in Community Question Answering[J]. IEICE Transactions on Information and Systems, 2009, 92-D(4): 736-739

[25] Sun K, Cao Y, Song X, et al. Learning to recommend questions based on user ratings[C]// Proc. of CIKM 2009. Hong Kong, China, 2009: 751-758

[26] Cao X, Cong G, Cui B, et al. The use of categorization information in language models for question retrieval[C]// CIKM 2009. Hong Kong, China, 2009: 265-274

[27] Cao Y, Duan H, Lin C Y, et al. Recommending Questions Using the MDL-Based Tree Cut Model[C]// Proc. of WWW 2008. Beijing, China, 2008: 81-90

[28] Bian J, Liu Y, Agichtein E, et al. Finding the Right Facts in the Crowd: Factoid Question Answering over Social Media[C]// Proc. of WWW 2008. Beijing, China, 2008: 467-476

[29] Suryanto M A, Lim E P, Sun A, et al. Quality-aware collaborative question answering: methods and evaluation[C]// Proc. of WSDM 2009. Barcelona, Spain, 2009: 142-151

[30] Tu X, Wang X J, Feng D, et al. Ranking community answers via analogical reasoning[C]// Proc. of WWW 2009. Madrid, Spain, 2009: 1227-1228

[31] **沈闻. 基于问答社区的个性化服务研究**[D]. **扬州: 扬州大学**, 2009

[32] Achananuparp P, Yang C C, Chen X. Using negative voting to diversify answers in non-factoid question answering[C]// Proc. of CIKM 2009. Hong Kong, China, 2009: 1681-1684

[33] Agichtein E, Liu Y, Bian J. Modeling Information Seeker Satisfaction in Community Question Answering[J]. ACM Transaction on Knowledge Discovery from Data, 2009, 3(2)

[34] Liu Y, Agichtein E. You've Got Answers: Towards Personalized Models for Predicting Success in Community Question Answering[C]// Proc. of ACL: HLT 2008. Columbus, OH, 2008: 97-100

[35] Guo J, Xu S, Bao S, et al. Tapping on the potential of q&a community by recommending answer providers[C]// Proc. of CIKM 2008. Napa Valley, California, USA, 2008: 921-930

[36] Qu M, Qiu G, He X, et al. Probabilistic question recommendation for question answering communities[C]// Proc. of WWW 2009. Madrid, Spain, 2009: 1229-1230

[37] Hu D, Gu S, Wang S, et al. Question recommendation for user-interactive question answering systems[C]// Proc. of ICUIMC 2008. Suwon, Korea, 2008: 39-44

[38] Zhou Y, Cong G, Cui B, et al. Routing Questions to the Right Users in Online Communities[C]// Proc. of ICDE 2009. Shanghai, China, 2009: 700-711

(**下转第** 54 **页**)

4.2.2 与 LEACH 的网络性能对比

由于 CCRP 采用了候选者的机制来平衡网络的成簇，使得网络分簇能更为平均，同时又限制了成为簇头的条件，避免了一些能量偏低的节点被选为簇头。其对分簇结构的优化更好地均摊了网络的能量消耗，从整体上延长了网络的生存时间。从图 6 中可以看到，同等轮数下 CCRP 的全网能耗小于 LEACH，整体生存时间得到延长。

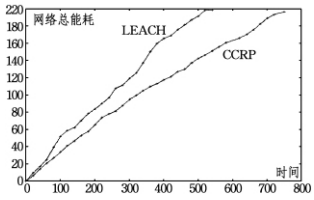

图 6 初始能量为 2J 时，LEACH 与 CCRP 网络总能耗对比

在数据传输阶段 CCRP 采用预测策略防止能量过低的节点继续担任高能耗的数据融合和传输工作，推迟了节点的死亡时间。并且由于簇头能及时发现失效的节点，将其剔除时间片序列，缩短了帧周期，增大了节点的可传输次数。从图 7 中可以看到，当网络进入整体能量已经偏低的生命期，CCRP 的第一个节点的死亡时间要比 LEACH 来得晚。

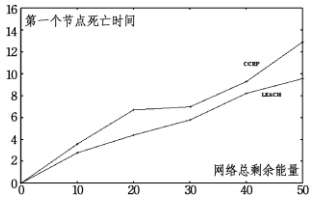

图 7 网络低能量状态时，LEACH 与 CCRP 的第一个节点死亡时间比较

结束语 在实际应用中，传感器网络的路由协议还有很大的优化空间和很多待解决的问题。对于能耗的苛刻限制使得要设计一个各方面均有优异表现的路由协议十分困难，而且设计的时候还要考虑到网络的其他层次和整体结构。本文对路由协议的研究都是在理想条件下进行的，现实中还要考虑到很多实际因素的影响。另外对于大型网络还可能需要进行多层分层，这也是今后研究的方向。

参 考 文 献

[1] Warneke B A, Pister K S J. MEMS for distributed wireless sensor networks[C]// The 9th IEEE International Conference on Electronics, Circuits and Systems. Piscataway, NJ, USA, 2002: 291-294

[2] Frodigh M, Johansson P, Larsson P. Wireless Ad-hoc networking: the art of networking without a network[J]. Ericsson Review, 2000, 4: 248-263

[3] 李建中，高宏. 无线传感器网络的研究进展[J]. 计算机研究与发展，2008，45(1)：1-15

[4] Akyildiz I F, Cayirci E, Su W, et al. Wireless sensor networks: a survey[J]. Computer Networks, 2003, 38(4): 393-422

[5] Heinzelman W, Chandrakasan A, Balakrishnan H. Energy-Efficient communication protocol for wireless microsensor networks [C]// Proc. of the 33rd Annual Hawaii Int'l Conf. on System Sciences. Maui: IEEE Computer Society, 2000: 3005-3014

[6] Heinzelman W. Application-Specific protocol architectures for wireless networks[D]. Boston: Massachusetts Institute of Technology, 2000

[7] Younis O, Fahmy S. Heed: A hybrid, energy- efficient, distributed clustering approach for ad-hoc sensor networks[J]. IEEE Trans. on Mobile Computing, 2004, 3(4): 660-669

[8] Zheng Zeng-wei, Wu Zhao-hui, Lin Huai-zhong, et al. CRAM: An Energy Efficient Routing Algorithm for Wireless Sensor Networks [C] // Proceedings: 19th ISCIS. Berlin, Germany: Springer-Verlag, 2004: 341-350

[9] 李成法，陈贵海，叶懋，等. 一种基于非均匀分簇的无线传感器网络路由协议[J]. 计算机学报，2007，30(1)：27-36

(上接第 23 页)

[39] Liu Y, Li S, Cao Y, et al. Understanding and Summarizing Answers in Community-based Question Answering Services[C]// Proc. of COLING 2008. Manchester, 2008

[40] Li B, Liu Y, Agichtein E. CoCQA: Co-Training Over Questions and Answers with an Application to Predicting Question Subjectivity Orientation[J]. Conference on Empirical Methods in Natural Language Processing. Honolulu, Hawaii, USA, 2008: 937-946

[41] Li B, Liu Y, Ram A, et al. Exploring Question Subjectivity Prediction in Community QA[C]// Proc. of SIGIR2008. Singapore, 2008: 735-736

[42] Cong G, Wang L, Lin C Y, et al. Finding question-answer pairs from online forums[C]// Proc. of SIGIR2008. Singapore, 2008: 467-474

[43] 高琦. 图书馆参考咨询与"百度知道"合作探析[J]. 图书馆学研究，2008(10)：90-93

[44] 杨思洛，毕艳娜. 搜索引擎的互动问答平台及其对数字参考咨询服务的启示[J]. 图书馆学研究，2007，51(2)：96-99

[45] 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

[46] Yeh T, Lee J, Darrell T. Photo-based Question Answering[C]// Proceedings of the 16th International Conference on Multimedia. Vancouver, British Columbia, Canada, 2008: 389-398

[47] Yeh T, Darrell T. Multimodal Question Answering for Mobile Devices[C]// Proc. of the 2008 International Conference on Intelligent User Interfaces. Gran Canaria, Canary Islands, Spain, 2008: 405-408

[48] Li G, Ming Z, Li H, et al. Video reference: question answering on YouTube[C] // ACM Multimedia 2009. Beijing, China, 2009: 773-776

[49] Hong L, Davison B D. A classification-based approach to question answering in discussion boards[C]// Proc. of SIGIR 2009. Boston, MA, USA, 2009: 171-178

[50] Livonia M. Quicken Loans Word of Mouth Marketing Award [EB/OL] https://www. quickenloans. com/about/press-room/quicken-wins-wommie-award, 2007-12-11

[51] Matt M. Part Two: Why Use Yahoo ! Answers[EB/OL]. http://www. smallbusinesssem. com/part-two-why-use-yahoo-answ-ers/1063/#ixzz0TgPSaIZZ, 2008-02-12