# 社区问答技术

蔡黎，周光有，赵军

（中国科学院自动化研究所模式识别国家重点实验室，北京 100190）

关键词：问答系统　问题检索　问题分类　质量评估

## 引言

随着互联网技术的普及和发展，网络已经成为人们日常生活中必不可少的部分。互联网极大地丰富了人们的日常生活，人们可以在网上购物、交友等，但是更多的时候人们是通过网络获取信息。随着 Web2.0 时代的到来，互联网正经历着一场巨大的变革，网络服务形式从单纯的信息发布开始向人人参与的方向发展。在人工智能、自然语言处理和信息检索技术的推动下，网上出现了大量基于 Web2.0 的新型服务模式，比如：论坛 (Forum)、博客 (Blog，如新浪博客等)、社区问答 (community-based QA，如 Yahoo! Answers、百度知道、腾讯搜搜问问、新浪爱问等)、社交网络 (social network based community，如校内网、Facebook 等)。社区问答是其中的一种典型应用，它的出现为互联网的知识分享提供了新的途径和平台，并为问答技术带来了新的生机。目前，社区问答已经逐渐受到业界同行和相关领域研究人员的重视，成为研究热点。传统的搜索引擎通过关键字匹配，返回大量网页供用户选择，给用户带来了很大不便。而社区问答是一个开放、互动的网络平台，通过用户参与，利用网络用户的集体智慧，提供问题的直接答案。目前主要的开放式社区问答平台有：

爱问(http://iask.sina.com.cn/)：是第一个中文社区问答系统，由新浪于 2004 年自主研发上线。爱问调动全体网民参与提问与回答，对于用户提问，系统将问题和回答信息有序地排列出来，供用户利用或参考，让用户彼此分享知识和经验。

百度知道(http://zhidao.baidu.com/)：是全球最大的中文搜索引擎百度推出的服务，它的功能和服务与新浪爱问大同小异，提供了一个涵盖范围包罗万象的网络在线互动问答平台。用户可以在“百度知道”中提交问题或者回答其他用户的问题，这些问题既可以是特定领域内的学术性或者知识性问题，也可以是日常的生活问题。目前“百度知道”的问题分类已经涉及电脑/网络、体育/运动、健康/医疗、生活/时尚、电子数码、教育/科学等二十三大领域，并且这些领域还在不断扩充。表 1 是“百度知道”的一个典型问题和答案的结构图。

| URL | http://zhidao.baidu.com/question/80273346.html | |
|---|---|---|
| 问题 | 分类 | 百度知道 > 电子数码 > 手机/通讯 > 手机使用 |
| | 标题 | 手机联想 P80 和三星 S3600 哪个好些？ |
| | 描述详细 | 如题，2 个价位都差不多，都 1000 出头。三星的那款好几项参数都不如联想的那款，如：屏幕大小、色彩、摄像头等，但是网上对联想手机的评价似乎不怎么样，而且做手机明显三星强于联想，我到底应该选择哪款呢？ |
| | 提问者 | jinzechen |
| | 时间 | 2009-1-4 15:44 |
| 最佳答案 | 答案内容 | 要我觉得，你选择 NOKIA 吧？联想，不好，质量什么的都觉得不尽人意！三星，更不好，TMD 我一直不喜欢三星，功能一般，贵得吓人，凭什么这样来宰我们？？质量还不见得好。现在国内推荐的，还是诺基亚和多普达！什么？贵？不啊，我觉得 NOKIA 不贵啊？现在谁还买行货啊，都买全新港货或者欧货了，谁还敢说贵？<br>兄弟，我以一名用机者来讲，这两个都别想了吧，换个诺基亚，扎实 ！<br>大吴手机技术工厂 |
| | 回答者 | weition2 |
| | 回答时间 | 2008-12-26 15:15 |

基金资助：本文受国家自然科学基金项目（60875041，61070106）资助

| | | |
|---|---|---|
| 其他答案 | 答案内容 | 我买过联想手机，那个郁闷。出的问题很折腾，但是可以用，就是收短信延迟已一天，诸如此类问题。。不过联想很适合中国人，至于韩国货。。<br>就算咯<br>就不能选诺基亚吗。 |
| | 回答者 | mlr06 |
| | 回答时间 | 2008-12-26 15:11 |

表 1 一个问答页面主要结构图

Yahoo! Answers(http://answers.yahoo.com/)：是美国雅虎公司推出的全球最大的英文问答社区。目前学术界大部分研究都是在 Yahoo! Answers 的数据集上开展的，它的组织结构和形式与“百度知道”类似。用户可以在“Yahoo! Answers”中提交问题或者回答其他用户的问题，问题的长短和领域不限。“Yahoo! Answers”的问题分类涉及到二十六个大类，如：健康 (Health)、运动 (Sports)、旅游 (Travel) 等等，1,262 个小类，分类层次结构如图 1 所示。

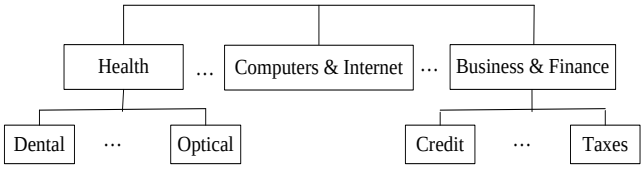


图 1 Yahoo! Answers 分类层次结构

某种意义上，社区问答可以看作一种新型的搜索引擎，更确切地说，是对已有搜索引擎功能的一个补充[1]。在这种服务体系下，用户可以根据自己的具体需求有针对性地提出问题，通过一系列积分奖励机制发动其他用户，来悬赏该问题的答案；同时，这些问题的答案又可以进一步作为搜索结果，提供给其他有相似需求的用户，从而达到知识分享的目的。借助这个网络平台，用户个人的隐性知识变成了显性知识，个体智慧变成了集体智慧，用户既是搜索引擎的使用者同时也是知识的创造者，充分体现了 Web2.0 以人为本的核心理念。

## 社区问答系统的构成

用户在互联网上应用社区问答的主要流程如图 2。用户首先提问，如果问答对历史记录(QA Archive）中有这个问题或相关问题的记录，则直接返回给用户；如果没有，则把问题放到网上，由其他互联网用户回答，然后由提问者选择最佳答案，直至最后关闭问题，同时将这个问答对保存到问答对历史记录中。

一个典型的社区问答通常由两个相对独立的部分构成，一部分用来维护用户信息，包括积分、等级、提问数、采纳率以及在问答社区中的活动等，称为用户管理系统；另外一部分则是问答社区的核心系统，主要包括问题分析、问题检索等主要功能。

用户是整个系统的主体，基于系统安全性考虑，用户只有在成为合法会员后才享有参与提问和回答的权利。为了验证合法会员，系统需要验证用户身份。用户在未登录的情况下，不能参与提问和回答，只能搜索答案、浏览问题、评价问题和参与投票。用户登录成功后，主要权限包括：提出问题、回答问题、处理自己的问题等。目前大多数问答社区都引入了个人中心的概念，例如在“百度知道”的个人中心中可以实现我的资料、我的问题、我的答案、我的积分、我的称号等功能。为了激励用户参与的积极性，问答社区中的合法会员都具有相应的头衔、等级机制。系统当前广泛采用的是积分制，用户在不同积分区间，享有不同称号。用户可以根据自己的知识水平回答自己感兴趣的问题来增加积分。

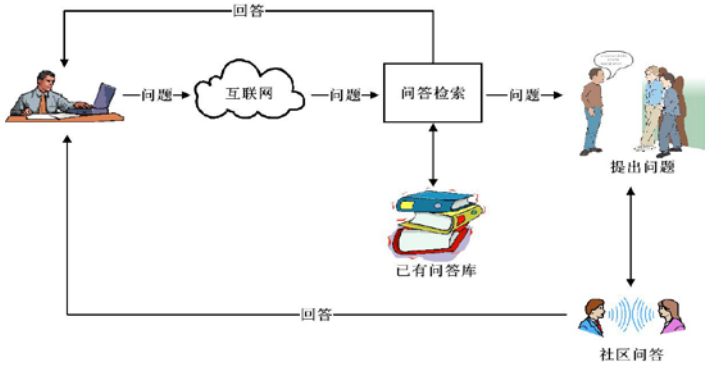


图 2 社区问答系统构成

问答是社区系统的核心部分，实现提问和回答两方面的功能，一般主要由问题分析、问题检索、答案处理三部分组成，如图 3 所示。

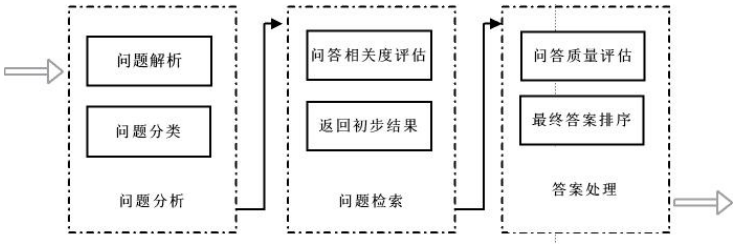


图 3　问答系统构成图

“问题分析”通过对问题语法、语义等进行解析，获取该问题的主题信息。“问题检索”从大量问答对历史记录中找到与用户相关或相近的问题，其中问题相关性计算是问题检索模块的关键所在。“答案处理”对检索出的答案进行排序以及质量评估，从而将最符合用户意图的答案返回给用户。

## 社区问答研究的关键技术

社区问答系统与传统问答系统（自动问答系统[9]、常见问题问答系统[10]等）有明显的区别。自动问答系统可以接受用户以自然语言形式的提问，并从海量的数据源中自动查找出问题的回答。由于受限于自然语言处理和人工智能技术的发展，目前自动问答只能接受相对简单的问句，问句形式的约束性很强，比如“毛泽东的生日？”等等。由于离真实用户的信息需求相差比较大，自动问答应用范围很小。常见问题问答系统是一种基于常问问题集的问答系统。常见问题问答系统数据集的质量很高，而且基本上是特定领域的，例如“电脑的常见问题的搜索”。相对于传统的问答系统，社区问答数据存在以下特点：数据规模大，数据质量噪声大，领域更宽泛。以下分析社区问答需要解决的关键技术。

1. 相关性问题检索

问答检索就是从数据集中找出能够解决用户提问问题的相应答案或问答对。
问答搜索通常采用相似问句匹配的方式，从问答对历史记录中搜索与问句语义相同或相似的已有问题，并将已有问题和答案返回用户。这种方式所面临的挑战是问句检索中的词汇鸿沟问题，也就是两个语义相同的问句可能采用的词语表述形式完全不相同。例如“如何减肥”和“怎样瘦身”，其实意思是一样的，但是两者的用词是完全不一样的。传统的信息检索模型，如向量空间模型、语言模型，都是基于词匹配的原理。因此利用传统的检索模型很难检索出语义相关但用词不一样的句子，但是这样的句子在社区问答中大量存在。

单语言翻译(monolingual translation)概率模型近年来被引入了信息检索框架中，通过 IBM 翻译模型[11]，从海量单语问答语料中获得同种语言中两个不同词语之间的语义转换概率，从而在一定程度上解决词汇鸿沟问题[3]。例如和“减肥”对应的概率高的相关词有“瘦身”、“跑步”、“饮食”、“健康”、“远动”等等。这方面的主要研究有：Joen[2] 和 Xue[3]把从问答对语料中训练出来的单语词与词之间的翻译概率作为词汇转移概率融入传统的基于语言模型的检索框架中，并由此计算出问题间的相似度，从而进行问题检索。实验证明该方法可以有效地解决词汇鸿沟的问题。

2. 问答对质量分析

社区问答是 Web2.0 技术的一个典型应用，而 Web2.0 的精髓就是用户参与网站内容创造，一方面 Web2.0 网站为用户提供了更多参与的机会；另一方面，也意味着相对 Web 1.0 的网站的内容，用户编辑的 Web2.0 内容的整体质量和稳定程度都差许多。例如，表 2 就是一个完全没有价值的问答对。
除此之外，社区问答对的数量也以惊人的速度增长。例如“百度知道”经过八年的发展，其数据已经超过 1 亿，保守估计每天以百万增长。面对如此海量，索引数据的质量而不是数量成为社区问答检索研究更加关注的问题。

| URL | http://zhidao.baidu.com/question/191988710.html?fr=qrl&cid=175&index=2 |
|---|---|
| 问句 | 这个地球上还有比我更帅的人吗？ |
| 答案 | 绝对没有。!!! |

表 2　一个低质量问答对的例子

自动分析和找出高质量问答对将对用户的满意程度产生重要影响。目前的人工智能和自然语言处理技

术还远远不能理解文本的语义内容。传统的网页质量分析是通过 PageRank[4]算法对每个网页进行评估等级，其基本思想是高质量的网页和高质量的网页有链接。但是与网页链接不一样，社区问答之间的链接实际是上用户和问题、用户和回答的链接。Agichtein[5]等假设问题的质量和提问用户的水平相关，答案的质量和回答问题的用户的水平相关。他们把问题质量和答案质量与提问用户和回答用户的水平挂钩，从而对问答对质量进行评估。如表 3 所示，这是一个社区问答用户在“百度知道”里的记录。系统可以根据用户在社区问答中的历史记录来衡量用户的水平，从而问题质量和答案质量进行评估。

| URL | http://passport.baidu.com/?business&aid=6&un=mlr06#2 | |
|---|---|---|
| 经验值 | 总分 | 102 |
| | 回答得分 | 822 |
| | 掌门积分 | 0 |
| | 知道处罚 | -280 |
| | … | … |
| 财富值 | 总分 | 10 |
| | 知道财富 | 80 |
| | … | … |
| 知道明细 | 回答数 | 119 |
| | 回答被采纳 | 1 |
| | 采纳率 | 17% |
| 回答过的历史问题 | 电脑上装有蓝牙，以前能用现在怎么不能用了？ | |
| | 旧电脑 能卖多少钱 | |
| | 帮忙挑个笔记本电脑~~ | |
| | … | |

表 3　一个用户页面主要结构图

3. 用户分析

Web2.0 的发展不仅改变了人类信息产生的方式，也给人类的信息传播方式产生了深刻的影响，使人类的信息环境发生了巨大变化。Web2.0 提供的多种网络服务，如博客(Blog)、社交网络(Myspace、FaceBook)、媒体共享(Flickr、YouTube)、社会化标注(Delicious)、协同创作(Wikipedia)等，为用户提供了新的信息交流和共享方式，将 Web 构建成为一种虚拟的社会空间。这种虚拟社会空间使信息的扩散真正突破了时空和地域的限制，从而对个人和社会都产生着重要影响。

社区问答是社会化媒体的一种形式，对社会化媒体的用户进行分析，是社区问答发展的重要途径。因此，如何从社区问答系统用户之间形成的社会化网络中，挖掘出网络用户行为之间的相互关联关系是社区问答系统面临的难题。专家查找(expert finding)是目前社区问答领域用户分析最有效的应用。对于一个用户提问，专家查找就像一个中介，找到那些能够解决用户需求并且愿意提供帮助的用户（称为专家）。例如：对于用户提问“中科院自动化所 2010 年的招生分数线是多少？”，社区问答系统所能够做的最好是，找到对考研非常熟悉而且对中科院自动化非常熟悉的用户，然后把这个问题推荐给这个专家用户。为了找到这个专家，社区问答系统所能利用的信息来源就是用户以前回答过的问题记录。如表 3 所示，最后一栏记录了一个社区问答用户在“百度知道”里回答问题的记录。社区问答系统通过分析用户所回答过的历史问题挖掘出用户的兴趣和专长。专家查找在社会计算的多个领域被广泛研究。Lin[6]研究了专家用户推荐的问题；Jurczyk[7]利用 HITS[8]算法基于用户之间的链接结构对用户权威性进行分析，从而找出社区问答系统中的权威用户，实验表明这种方法效果显著。

# 结语与展望

社区问答系统是一种新兴的社会化网络信息服务系统，对人们快速准确地获取信息和方便地信息交流提供了很大帮助。近年来，社区问答已逐渐成为互联网领域和数据挖掘领域的一个研究热点。社区问答未来的主要发展方向有两个：（1）提供优质的问题答案搜索服务：相对于传统的基于查询词的信息检索，社区问答检索的信息需求更加明确。未来的发展方向是充分利用社区问答提供的各种信息，融合自然语言处理技术，如问题聚类、句法分析、语义角色标注、信息抽取等，通过对问题语义进行更深入的挖掘，为用户提供更优质的问题答案搜索服务。（2）增加社区问答的互动性：相对于博客、社交网络等网络社交平台，社区问答的互动性相对比较弱，建立提问用户和回答用户之间的联系，使他们之间有更多互动，增加用户和社区问答之间的黏度，才能更好地扩大知识共享。总体来说，社区问答系统有广阔的发展空间，社区问答系统将会结合网络搜索和社会计算技术，成为社会搜索和知识库系统构建的重要组成部分。

**参考文献**

[1] 高兵，问答式社区的标签推荐技术研究。哈尔滨工业大学硕士论文，2009 年 6 月.
[2] J. Jeon, W. B. Croft, and J. Lee. Finding Similar Questions in Large Question and Answer Archives. In proceedings of CIKM'05, pages 84-90, 2005
[3] X. B. Xue, J. Jeon, and W. B. Croft. Retrieval Models for Question and Answer Archives. In proceedings of SIGIR'08, pages 475-482, 2008
[4] S. Brin and L. Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. Computer Networks and ISDN Systems, Volume 30, Issue 1-7, pages 107-117, 1998
[5] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G.Mishne. Finding High-Quality Content in Social Media. In proceedings of WDSM'08, pages 183-194, 200
[6] J. M. Kleinberg. Authoritative Sources in a Hyperlinked Environment. Journal of the ACM, 46(5), 1999
[7] P. Jurczyk and E. Agichtein. Discovering Authorities in Question Answer Communities by Using Link Analysis. In proceedings of CIKM'07, pages 919-922, 2007
[8] J. M. Kleinberg. Authoritative Sources in a Hyperlinked Environment. Journal of the ACM, 46(5), 1999
[9] H. T. Dang, J. Lin, and D. Kelly. Overview of the TREC 2006 Question Answering Track, 2006
[10] R. D. Burke, K. J. Hammond, V. A. Kulyukin, S. L. Lytinen, N. Tomuro, and S. Schoenberg. Question Answering from Frequently Asked Question Files: Experiences with the FAQ Finder System. Technical report, University of Chicago, 1997
[11] P. F. Brown, V. J. D. Pietra, S. A. D. Pietra, and R. L. Mercer. The Mathematics of Statistical Machine Translation: Parameter Estimation. Computational Lingustics, 1993