中文社区问答中问题答案质量评价和预测

李 晨 巢文涵 陈小明 李舟军
(北京航空航天大学计算机学院 北京 100191)

摘 要 知识共享型网站为自动问答系统带来了新的研究契机。但用户提供的问题及其答案质量参差不齐,在提供有用信息的同时可能包含各种无关甚至恶意的信息。对此类信息进行判别和过滤,并选取高质量的问题与答案对,有助于在基于社区的自动问答系统中重用相关问题的答案以提高问答系统的服务质量。首先从中文社区问答网站上抓取大量问题及答案,利用社会网络的方法对提问者和回答者的互动关系及特点进行了统计与分析。然后基于给定的问答质量判定标准,对3000多个问题及其答案进行了人工标注。并通过提取文本和非文本两类特征集,利用机器学习算法设计和实现了基于特征集的问答质量分类器。试验结果表明其精度和召回率均在70%以上。最后分析了影响社区网络中问答质量的主要因素。

关键词 社区问答,社会网络,机器学习,问题答案质量评价和预测,人工标注
中图法分类号 TP391.1 **文献标识码** A

Quality Evaluation and Prediction for Question and Answer in Chinese Community Question Answering

LI Chen CHAO Wen-han CHEN Xiao-ming LI Zhou-jun
(School of Computer Science, Beihang University, Beijing 100191, China)

Abstract The rise of Knowledge-sharing platform on the Internet in China provides a new approach for Automatic Question Answering. However, the quality of User-Generated Content in such social networks may vary significantly, from useless information to malice spam. Identifying and filtering such content are particularly important to improve users' experience and the performance of Question Answering System. We first extracted a set of question answer content from Chinese Community Question Answering site, investigated a series of statistic characteristics on the interaction of participants, and then manually annotated quality of a subset of these questions and answers. By combining text features and non-text features provided by the community extracted from those questions and answers, we established a content quality classification model for evaluation and prediction. We find that this model is able to distinguish high-quality ones from others with considerable accuracy.

Keywords Community question answering, Social networks, Machine learning, Question and answer quality evaluation and prediction, Human annotation

1 引言

在目前互联网上非常流行的知识共享型网站中,每天都积累成千上万的问答对,累计已达上亿级别,如百度知道、Yahoo! Answer等。这些问答对能够覆盖用户所关心的日常工作学习中最常见的问题,并且问题和答案文字都没有上下文假设,因此可以抽取上述问答对建立自动问答知识库。基于问答对库的问答系统中,答案是人工给出的。对于用户输入的问题,可以快速地检索到答案,而答案的质量仅取决于问答对库的质量。所以,自动问答系统中知识库的搭建对后续的相似问题判断、答案抽取等有着至关重要的作用。然而,在社区问答中,社区内容是由用户产生,不可避免地存在着大量的无关和垃圾信息。例如,

低质量的答案:

提问:怎样可以最有效地瘦臀?

回答:蹭树。

提问:为什么我玩3D游戏时会头晕?

回答一:小脑不发达;

回答二:大脑不发达;

回答三:大小脑都不发达。

低质量的问题,提问者或许仅仅是为了吸引大量的回帖,而并不是希望解决实际的问题:

提问:一同学被蜘蛛咬了,问我会不会变蜘蛛侠?我是认真的哦,大家好好回答我哦。他正着急着呢!

到稿日期:2010-07-11 返修日期:2010-10-14 本文受国家自然科学基金项目(90718017)和教育部高等学校博士学科点专项基金(20070006055)资助。

李 晨(1985—),男,硕士生,主要研究方向为社区问答、机器学习、数据挖掘等,E-mail:lichen782@yahoo.com.cn;巢文涵(1979—),男,博士,讲师,主要研究方向为机器翻译、自然语言处理;陈小明(1980—),男,博士生,主要研究方向为自动问答系统、自然语言处理;李舟军(1963—),男,博士,教授,博士生导师,主要研究方向为高可信技术、安全协议分析、数据挖掘与信息检索。

提问：如何练成天马流星拳？

提问：啦啦啦，测试下～～～

上述问题和答案频繁出现在社区问答中，严重地影响了问答系统的用户体验，同时也降低了答案抽取的精度。如何判别和过滤无关和恶意的信息，选取高质量的问题与答案对，有助于基于社区问答知识库的自动问答系统重用相关问题的答案，提高自动问答系统的服务质量。

为了能够建立高质量的问答对知识库，从社区问答平台中提取问答对时必须使用有效的问答对质量评估方式，对问题答案对的质量进行评估。为解决上述问题，本文利用问题答案对在社区问答平台中的文本特征和非文本特征，进行逻辑回归分类训练，从而建立有效的分类器，对问答对质量进行评估。文本特征中主要包含文本视觉特征（例如标点符号密度、平均词长、文本熵等）和文本内容特征（例如文本内容词比例、疑问词密度、相关词覆盖等），并提取中文自动差错广泛采用的特征（例如单字密度等）；非文本特征包含用户的权威指标、答案问题状态等面向网站结构的特征。我们利用训练得到了的回归模型对问题和答案的质量进行评估，得到了良好的结果，其中精度和召回率在70%以上。

本文第2节介绍国内外相关的研究工作；第3节描述本文的数据来源和初步统计的社会网络特点；第4节给出人工标注结果和本文建立的答案问题质量分类模型；第5节给出分类模型实验结果并讨论；最后总结全文并给出未来的工作计划。

2 相关工作

对互联网上内容质量进行评价的工作最早可以追溯到Google的PageRank[1]算法，以及其他早期使用链接分析[2]的方法。Zhou和Croft[3]把检索文档的信噪比和KL距离作为文档质量的判断指标，并使用信噪比和KL距离的合成分值作为似然检索模型中文档的先验概率。但上述方法所解决的是网页文本的质量评估，并不适用于短文本和不规范用语较多的社区问答对质量评估。

在判断答案质量上，AnswerBus[9]使用问题类型和答案的匹配、命名实体识别、指代消解、冗余度、词匹配等指标来为候选答案排序。E. Agichtein和Y. Liu使用原有用户交互的历史信息作为分类器的特征输入，来判断提问者是否得到满意的答案[10]。Sofia J. Athenikos和Hyoil Han等[11]使用描述逻辑来表示知识，通过逻辑推理来归纳高质量的答案。Tu Xudong等[12]通过提取Yahoo! Answer的问题和最佳答案对之间的关系来运用类比推理的方法为新问题找到最佳答案。

而问题的质量判断主要集中在如何从大量文本中提取问题句子。V. Jijkoun提出URL中大多数包含"faq"字符串的是"常见问答"列表的网站，通过构造URL挖掘有效问题[13]。Y. S. Lai和K. A Fung采用列表探测（List Detection）的方法来提取问题和答案对[14]。G. Cong等利用频繁项集挖掘算法从论坛上提取问题-答案对，包括对问题的识别和在同一主题中对答案的识别[15]。但基于严格句法标注的序列挖掘的前提是文本集是高质量的、难以应用于社区网络中的文本内容。

社区问答是社会网络的一类，因此判断问题答案的质量既有其自身的特点，也与社会网络中用户产生内容的质量评价具有一致性。用于社会网络中的链接分析[8]方法都适用于社区问答中用户权威度分析[4-6]。Zhang Jun等[6]认为上述链接分析法在聚合度较低的网络中表现比较好。J. Jeon等[7]的工作表明，当前用户的级别（采纳率）与该用户给出的答案质量评分有较大的相关性。

社区问答（CQA）内容质量评估方法是目前国外的研究热点[4-7,10,16]。社区问答系统中判断用户产生内容（User-Generated Content）的质量主要依据社区统计信息（用户积分、答案票数、点击次数等）[7,17]、启发式的文本特征或前两者之间的结合值[18,20,21]。其中，在文本特征方面主要借用了论文自动评分[19]的方法，对文本的可读性、连续性和拼写与语法错误等进行检查。文献[21]发现人工标注的最佳答案往往不是提问者采纳的答案，但不同人工标注者之间对相同数据集的评分有很高的相关性。

而在中文社区问答中还没有看到相关的工作。国内的自动问答研究主要集中于对问题分析、答案抽取和本体研究上[22,23]。主要原因是中文相对于英文在表现形式、语法研究和资源收集等方面有很大不同。

3 数据收集和初步统计

3.1 数据收集

百度知道是目前全球最大的中文互动问答平台，该平台还通过奖惩机制来激励用户回答问题赚取积分。根据百度知道的网站介绍，用户提出问题和回答问题的流程如图1所示。

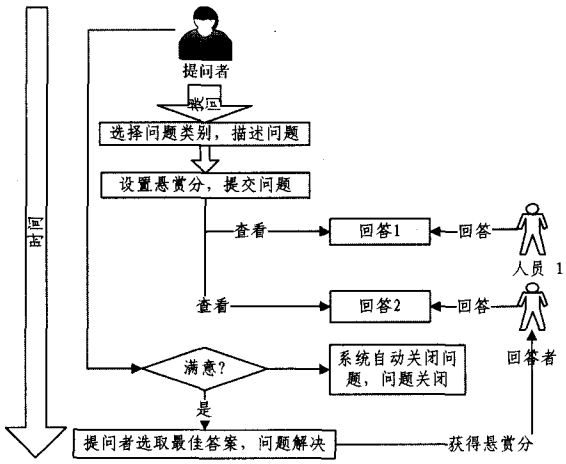
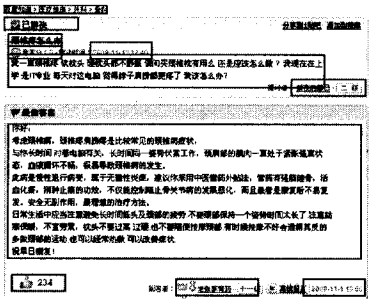


图1 百度知道参与流程

我们从百度知道上收集了从2005年6月12日到2005年7月12日的33637个问题以及与之对应的答案145184个，平均每个问题4个到5个答案。将问题和答案相关的属性（见图2）提取出来，并存储到关系型数据库中。根据提取的信息，得到27964个用户和356个分类，其中前20个分类中的问题数目占了55%。通过对提取的信息的分析可以看出用户所关注的话题主要分布在计算机网络、互联网和个人感情方面。

在所有问题中，11197个问题的状态是"已解决"，占解决问题数目的33.3%；剩下66.7%的问题状态是"已关闭"。可见绝大部分用户没有及时地处理问题，或对答案质量本身不满意。根据问题所对应的答案数量分析，有12.6%的问题没有任何答案，54.9%的问题拥有1到4个回答。基于上述提取的信息，下节将对该社会网络进行初步的统计。



红框代表非文本的信息；蓝框代表文本部分，
也即标题、问题描述、答案内容。

图 2　百度知道页面信息

3.2　初步统计

本文利用有向图对用户之间的相互关系进行建模，把社区网络中用户之间的相互关系映射成一个有向图 $G=(V, E)$。其中，V 是节点的集合，代表社区问答的参与者；E 代表边的集合，是用户之间的相互关系。在社区网络中，如果用户 u 回答了用户 v 的一个问题，就有一条从 v 到 u 的边。因此，一个节点的入度(In-Degree)就是它回答问题的数目，出度(Out-Degree)则代表了它提出问题的数目。本节从该社会网络的领结结构、度分布和度相关等角度进行了统计，并做了链接分析算法的比较。

3.2.1　领结结构

领结结构(bow tie structure)用于表征某个网络中节点之间关系类型所占比例的一种统计，分为中央(Core)、入部(In)、出部(Out)、卷须(tendrils)以及管部(tubes)。最初 Broder 等发现互联网各网站链接关系呈现出标准的对称领结结构[25]。而在社区问答中[6]，中央部是该有向图的强连通子图。中央部的任意一个用户沿着答案-问题链接可以到达该部的任意一个其他用户，体现了用户频繁的互相帮助关系；入部的用户是那些只提问题的用户；出部的用户仅回答中央部用户的问题；卷须和管部的用户则是那些仅仅链接出部和入部的用户，但不和中央部用户打交道，这部分用户仅回答入部用户的问题或者仅向出部的用户提出问题。

表 1 列出了互联网[25]、Java Forum[6]和百度知道的领结结构对比。

表 1　互联网、Java Forum 和百度知道的领结结构对比

	Core	In	Out	Tendrils	Disconnect
Web	27.7%	21.2%	21.2%	21.5%	8.0%
Forum	12.3%	54.9%	13.0%	17.5%	1.9%
Zhidao[1]	20.8%	35.7%	27.0%	15.9%	0.7%
Zhidao[2]	0.3%	53.9%	20.1%	16.4%	6.5%

可以看到早期(2005 年 6 月)百度知道的结构和互联网网页链接结构非常相似。这表明在初期百度知道中有大约 1/5 的用户积极地相互回答和提出问题。而在后期，百度知道并没有这样的领结结构，其核心部分所占比例不超过 0.3%。后期的数据表明，多数用户登录社区只提出自己的问题，而不再热衷于相互回答和提出问题。

3.2.2　度分布

度分布广泛用于研究在线社区中用户的相互关系。图 3 和图 4 分别给出用户回答问题数和提出问题数的概率分布，虚线是拟合曲线。

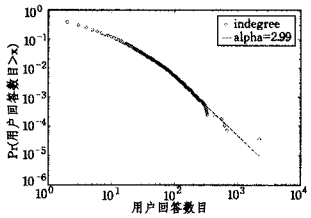


图 3　用户回答问题数的概率积累曲线

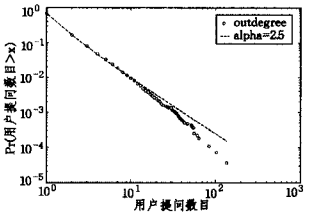
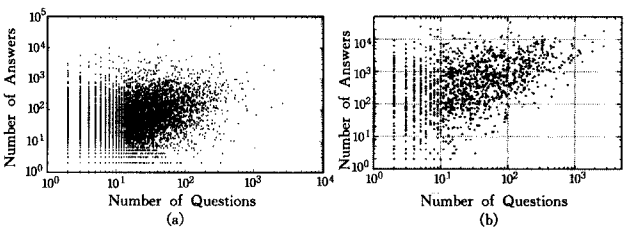


图 4　用户提出问题数的概率积累曲线

指数分布参数分别为 2.99 和 2.5，造成概率积累曲线下降的速度很快。

事实上，仅提出了一个问题的用户占到 52.3%；同时，没有贡献任何答案的用户占到总用户的 37.7%。这表明相当多的用户并不是社区积极参与者。图 5 是用户提出问题和回答问题的散点图，其中每一个点代表一名用户。

与文献[20]一样，存在一大群活跃度非常高的群体。但从互动特点方面来看，百度知道与文献[20]中研究的 yahoo answer 还是有明显的区别。yahoo answer 的散点靠近 $x=y$ 轴分布，并偏上方：靠近 $x=y$ 轴分布说明用户都积极地提出问题并积极地回答别人的问题；偏上方说明用户一般会积极回答但消极提问。而百度知道中，散点分布靠近 $x=y$ 的下方，意味着用户积极提问但消极回答。

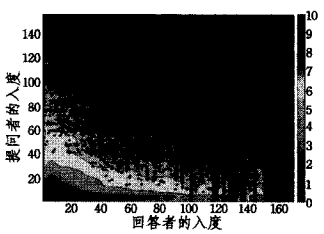


横轴代表用户提出的问题数目，纵轴代表用户回答数目；
(a)是百度知道的散点图，(b)是文献[20]的散点图。

图 5　用户提出问题和回答问题的散点图

3.2.3　度相关

我们给出每对提问者-回答者的入度(即回答问题数目)的简化图[6]，如图 6 所示。这个图颜色代表这样的点的个数有多少，颜色越冷，这样的点越少。比如回答者入度 20(x 轴) vs 提问者入度 10(y 轴)，取对数是对值取对数。



图中每点代表具有这样用户入度对频数

图 6　百度知道回答者和提问者入度相关性色彩图

[1] 百度知道在 2005 年 6 月到 7 月的数据，这时候百度知道刚刚内测完毕，进行公测，具有高比例的中央部分，当时的积分优惠与内部测试人员参与有关。

[2] 百度知道在 2006 年 11 月到 12 月的数据。

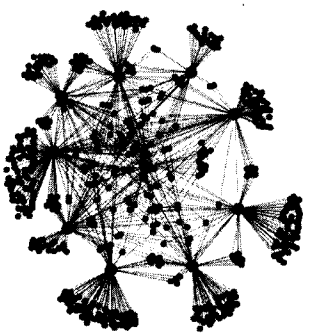
社会网络的相配性(assortativity)是指网络中度大的节点倾向于与度大的节点相连。如果一个用户回答问题的数目能够衡量该用户的权威性,那么社区问答中的正相配性就是指相同权威的用户之间往往互相回答问题。

使用 Pearson 相关系数来衡量提问者和回答者入度之间的相关性。Pearson 相关系数定义如下:

$$p=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{1}$$

式中,x 表示提问者的入度,y 表示回答者的入度。百度知道中并没有明显的相配性,提问者和回答者对之间入度的 Pearson 相关系数为−0.013,其表明几乎不具有相关性。

从图 6 可以看到,同等权威的用户都会有相同数量用户给予他们帮助,而且这些用户来自于不同的权威层次。图 7 给出更直观的可视化网络,这里展示的是入度前 10 名用户的网络拓扑结构。注意到他们的邻居既包括即时提问者(外围),也包括积极参与者(内部)。



他们的邻居既包括即时提问者(外围),也包括积极参与者(内部)。

图 7　入度前 10 名(红点)用户的网络拓扑结构

3.3　链接分析实验

鉴于用户的权威水平与其产生的社会网络内容之间是相关的,本文采用链接分析算法 ExpertiseRank[6] 和 HITS[4] 对节点的权威值进行排序。如果用户 A 回答了用户 U_1,U_2,U_3,…,U_n 的问题,那么用户 A 的 ExpertiseRank 得分为:

$$ER(A)=(1-d)+d(ER(U_1)/C(U_1)+\cdots+ER(U_n)/C(U_n)) \tag{2}$$

式中,$C(U_1)$表示节点 U_1 的出度。这里的 d 是松弛因子,一般取 0.8。

HITS 对每个节点赋予两个值:hub 得分和 authority 得分。每次迭代更新所有节点的两个值。

在本文中,用户所获得的好评数和采纳率是用户标准权威值。用用户回答问题数、ExpertiseRank 和 HITS 的得分与标准权威值进行 Pearson 相关性检验,结果如图 8 所示。回答问题数与标准权威值的相关性要远远高于链接分析算法。

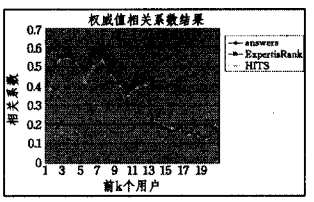


图 8　前 K 个用户回答问题数、ExpertiseRank 和 HITS 的得分与标准权威值进行 Pearson 相关性检验

由式(2)得知,容易对边进行加权。考虑对不同类型的链接分配不同权重。如:

$$ER(A)=(1-d)+d(w_iER(U_1)/C(U_1)+\cdots+w_jER(U_n)/C(U_n)) \tag{3}$$

式中,w_i 是:

1)0.5,如果用户 v 回答了用户 u 的一个问题;

2)$B+\log(\text{voteNum})$,如果用户 v 的回答被用户 u 采纳为最佳回答,其中 B 为参数,实验设置为 5;

3)n,如果用户 v 回答了用户 u 的 n 个问题。

加入权重后实验结果如图 9 所示。

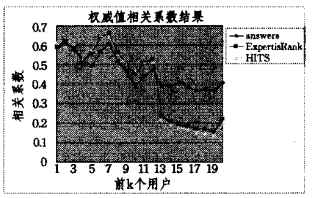


图 9　前 K 个用户回答问题数、ExpertiseRank 和 HITS 的边加权得分与标准权威值进行 Pearson 相关性检验

加权后比标准权威值相关系数有微小的提高。无论是否加权,链接分析方法得到的结果与利用回答问题数相比都显得不够稳定。原因首先在于我们选择的标准权威值不够精确,其次是用户之间的交互程度不如互联网网页链接密集,即 2.2.1 节中提到领结结构中的中央部分偏小;再次,链接分析算法对低匹配性网络效果不好[5],而根据 3.2.3 节的结果,社区问答的用户关系正属于这类网络。

综上所述,百度知道在 2005 年 6 月的社会网络具有较大数目的中央部(积极参与的用户);低相配性(用户倾向回答各个不同权威层次的提问);和专业求助论坛(Java Forum)相比,用户积极提问但消极回答;度分布上两极分化严重,即存在少数回答了上百问题的"超级用户",也存在大量仅仅提出一个问题的用户;基于链接分析得到的用户权威值并不稳定。

4　人工标注和质量特征

4.1　人工标注

两组标注者对抓取的数据进行了标注,并提供了标注标准,标注者在进行标注时可参考这些标准。具体标准内容如下。

问题:

1)用语是否规范,文字使用得体,阅读通顺连贯;

2)问题背景交待是否清楚和具体,描述内容可靠,不含歧义;

3)不存在恶意或者广告内容。

答案:

1)用语是否规范,文字使用得体,阅读通顺连贯;

2)是否和问题提出的内容相关,详细并能解决用户问题,具有针对性;

3)不存在恶意或者广告内容。

C. Shah[21] 认为人工标注者很难推断某个答案是否是"最佳答案",因此把提问者选择的最佳答案作为高质量的答案。然而问答社区,会出现如下情况:1)用户选择的最佳答案是错误的;2)用户没有选择最佳答案。所以直接把提问者选择的最佳答案作为高质量答案是不可取的,需要进行人工标注处理。

两组标注者分别对同一个问题集和对应的答案集进行标注，包括 3807 个问题和 17085 个答案。这些问题和答案被分成高(3 分)、中(2 分)和差(1 分) 3 个等级，标注者按照给出的标准进行标注。

图 10 标注结果显示低质量问题和答案只占少数。

表 2 给出两组标注者在问题标注上的混淆矩阵。标注者之间的 Kappa 相关系数为 0.31，说明在质量判定上存在一定分歧。

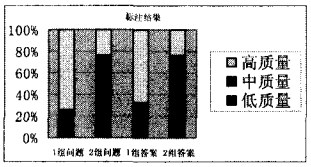


图 10 人工标注结果(1 组和 2 组)

表 2 两组标注者标注问题结果的混淆矩阵

1/2	低	中	高
低	78	76	129
中	154	437	1748
高	21	70	483

如果把中高质量归为一类，Kappa 系数上升到 0.88，则在答案方面也有类似的现象。

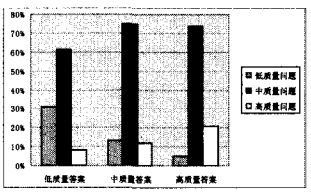


图 11 组 1 中答案质量和问题质量的关系

图 11 表明低质量的问题中往往出现低质量的答案，而较高质量的问题会吸引较高质量的答案。

在那些两组都标注为高质量的答案中，仅有 13.6%的答案被提问者采纳为最佳答案。所以仅以是否被采纳为最佳答案来判断高质量答案是不合理的。

4.2 特征提取

从百度知道页面上提取的问题和答案实体包含文本和非本文的信息，希望通过提取这些信息作为特征集合，训练分类器，用于判断它们的质量。

4.2.1 问题质量特征

参考图 2 的页面，我们列举提取的部分特征。

1)文本特征

√文本总长度：包括问题题目和描述的长度；

√内容词密度：问题题目和描述中名词、限定性副词、代词、动词等实词的比例；

√标点符号比重：大量使用诸如省略号或者“表情”，一般认为文本质量不高；

√最长单字散串：根据文献[24]中的实验，错误的用法往往导致分词后出现多个连续单字串；

√文本熵：包括基于词和字符的熵值。文本熵值往往用于检验文本的多样性。对于一段 λ 长度的文本，包含 n 个不同单词的文本熵定义如下：

$$E_T(p_1,\cdots,p_n)=\frac{1}{\lambda}\sum_{i=1}^{n}p_i[\log_{10}(\lambda)-\log_{10}(p_i)] \quad (4)$$

式中，$p_i, i=1,2,\cdots,n$ 是单词 i 在文本中的频率；

√疑问词所占比重；

√单字比重：分词后单个字在整个文本中所占比重；

√平均句子长度；

√平均词长；

√问题文本与背景文档集的 KL 差异(K-L divergence)[3]。直观上，高质量的文档和背景文档(如百度百科、维基百科等)在词分布上应该近似，本文选取百度百科[3]。文档 D 和背景文档 C 的 K-L 差异定义如下：

$$\text{KL-divergence}=\sum_{w}p_{coll}(w|C)\log\frac{p_{coll}(w|C)}{p_{doc}(w|D)} \quad (5)$$

式中，$p_{doc}(w|D)=\frac{\#Count(w,D)}{\|D\|}$，$p_{coll}(w|C)=\frac{\#Count(w,C)}{\|C\|}$。

2)非文本特征

·问题属性

√问题悬赏分：提问者可以在提问后设置悬赏分，用于鼓励其他用户回答；

√问题状态：已关闭或者已解决；

√问题得到回答的数目。

·用户属性

√回答数目：由 3.3 节得知回答数目是对用户权威比较稳定的估计；

√ ExpertiseRank 得分和 HITS_authority 链接分析得分：3.3 节中虽然它们在估计权威值上贡献不大，但考虑到其对辨别排名靠前用户的敏感，我们依然把它们作为特征；

√提问数目；

√积分[4]；

√采纳率：回答中有多少被采纳为最佳答案；

√好评率：该用户提供的答案得到的好评数；

√问题解决数目：提出的问题中有多少问题的状态是“已解决”；

√提问数目和答案数目的差值。

4.2.2 答案质量特征

1)文本特征

答案质量特征集和问题部分类似，但不包括疑问词所占比重，同时我们增加下列特征：

√内容词覆盖率：计算问题中的内容词(名词、限定性副词、形容词和动词等)在答案文本中重复的比例；

√类别距离：高质量的答案应该和问题从属于同一类别，例如医疗健康类问题中如果出现娱乐类别的答案，则不可能是较高的质量。类别距离的计算采用朴素贝叶斯算法，以百度知道上的一级分类的所有文档作为训练语料，得到答案文本向量和问题所属的一级类别向量的距离得分。文档 d 和类别 C_w 的距离定义如下：

$$Dist(d,C_w)=\frac{1}{P(d|C_w)} \quad (6)$$

$$P(d|C_w)=P(d)P(C_w|d)$$

2)非文本特征

[3] http://baike.baidu.com/

[4] 以下 3 项均通过百度知道用户信息中心获取，http://passport.baidu.com/

·答案属性

√是否被采纳:该答案是否被选为最佳答案;

√投票数:答案获得的票数;

√提问者评论:提问者有时会对答案做出评价。

·问题属性

即 4.2.1 节问题质量特征中的问题属性。

·用户属性

即 4.2.1 节问题质量特征中的用户属性。这里的用户指回答者。

5 分类实验和评价

5.1 问题质量判别

为避免两组标注者不一致造成的误差,我们提取两组标注结果的交集作为数据集,并去掉回答者或者提问者是匿名的情况,即:

$$U=\{x|\text{Annotat1}(x)=\text{Annotat2}(x)\} \quad (7)$$

得到标注一致的问题 1121 个,其中高、中、低质量百分比分别是 46.0%,32.7%,30.3%,然后利用逻辑回归建立分类模型。表 3 和表 4 显示对低质量和高质量问题的实验结果。

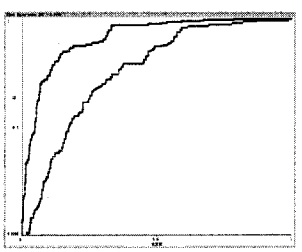
表 3 低质量问题(判别的精度 P、召回率 R、F-Measure 和 ROC 曲线面积 AUC)

特征集	P	R	F-Measure	AUC
文本	0.696	0.821	0.753	0.877
非文本	0.498	0.76	0.602	0.77
文本+非文本	**0.737**	**0.818**	**0.775**	**0.913**

表 4 高质量问题(判别的精度 P、召回率 R、F-Measure 和 ROC 曲线面积 AUC)

特征集	P	R	F-Measure	AUC
文本	0.673	0.621	0.646	0.802
非文本	0.563	0.504	0.532	0.716
文本+非文本	**0.723**	**0.764**	**0.743**	**0.833**

文本特征对判定低质量问题有一定帮助,这主要用于区分过于简短、描述不清或者包含大量标点和表情以及非正常中文字符的问题,这些问题能够很容易地通过文本特征过滤。文本特征在判别中比非文本特征要显著,原因是对问题来说,可用特征较少,某些重要的非文本特征,例如页面展示(PV)和点击率(CTR)无法取得。图 12 给出问题质量(低)判别的 ROC 曲线。



上方是文本+非文本特征,下方是非文本特征。

图 12 问题质量(低)判别的 ROC 曲线

5.2 答案质量判别

答案质量判别和问题部分类似。数据集中共包含 6628 个答案。其中高、中、低质量百分比分别是 39.6%,47.7%,12.7%。表 5 和表 6 显示对低质量和高质量问题分类的实验结果。

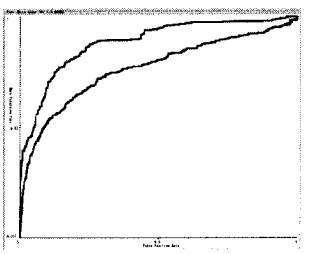
表 5 低质量答案(判别的精度 P、召回率 R、F-Measure 和 ROC 曲线面积 AUC

特征集	P	R	F-Measure	AUC
文本	0.593	0.617	0.605	0.835
非文本	0.612	0.581	0.596	0.849
文本+非文本	**0.658**	**0.642**	**0.65**	**0.849**

表 6 高质量答案(判别的精度 P、召回率 R、F-Measure 和 ROC 曲线面积 AUC)

特征集	P	R	F-Measure	AUC
文本	0.705	0.575	0.633	0.792
非文本	0.732	0.609	0.655	0.83
文本+非文本	**0.784**	**0.691**	**0.734**	**0.89**

分类模型对低质量答案的判别性能远远低于高质量判别,原因之一是低质量的答案实际上并不具有统一的特点,而高质量的答案在文本和非文本特征上都非常类似,例如拥有高权威值的用户、规范的文本表达等。但文本和非文本特征的配合使用确实能提高分类模型的性能。其次是数据集中的低质量答案数目有限,类分布不均匀。图 13 给出答案质量(低)判别的 ROC 曲线。



上方是文本+非文本特征,下方是非文本特征。

图 13 答案质量(低)判别的 ROC 曲线

表 7 和表 8 分别按照回归系数给出问题判别中前 10 项最影响质量判断的特征。问题判别中,文本特征起的作用往往大于在答案中的作用,但涉及自然语言的特征(类别距离和 K-L divergence)和问题与答案质量的相关性时却很低,原因是往往问题和答案都是短文本,不适合传统自然语言处理的长文本特点;再次是社区问答文本中出现的大量网络用语和文档集中词分布差异过大:我们观察到社区中用户用语习惯和百科全书有明显不同,而如果进行严格的语法查错,将极大地缩小高质量内容的比例。链接分析得分(ExpertiseRank 得分和 HITS 得分)都没有进入前 10 个显著特征,再一次证明它们并不适合该网络。

表 7 问题判别中前 10 个显著特征

序号	问题(好)	问题(差)
1	文本熵(字符)	文本熵(字符)
2	问题悬赏分	内容词比重
3	文本长度	文本长度
4	句子个数	最长单字散串
5	最长单字散串	句子个数
6	平均句子长度	提问者积分
7	词密度	提问者提问数
8	提问者回答数	问题悬赏分
9	提问者提问数	平均句子长
10	问题获得答案数	提问者回答数

表 8 答案判别中前 10 个显著特征

序号	答案(差)	答案(好)
1	文本长度	平均句子长度
2	句子个数	内容词比重
3	平均句子长度	文本长度
4	内容词比重	最长单字散串
5	回答者回答数	词覆盖率
6	回答者积分	回答者积分
7	文本熵	回答者提问数
8	最长单字散串	问题悬赏分
9	词覆盖率	平均句子长
10	问题悬赏分	回答者回答数

结束语 首先对中文社区的社会网络特点做了描述,发现百度知道在不同时期呈现出领结和非领结的交互结构,对用户度分布的统计表明社区中用户倾向积极提问但消极回答,整个网络呈现低相配性的特点;其次使用链接分析对参与回答的用户的权威进行评估,发现(加权)链接分析方法对估计排名靠前的用户的权威有较好的性能,但并不能对整个社会网络中的参与用户做稳定的权威估计,这是因为用于互联网 Web 页面引用的 PageRank 和 HITS 适用于节点之间频繁交互的网络结构;我们组织了标注者对部分数据集的质量进行人工标注,并选择了一系列文本和非文本(基于页面信息)的特征用于分类,取得了较好的效果。

下一步工作将把内容质量的得分引入到自动问答系统中,通过提升高质量问题和答案的排名来提高问答系统的服务质量和用户满意度。

参 考 文 献

[1] Brin S, Page L. The anatomy of a large-scale hypertextual Web search engine[J]. Computer Networks and ISDN Systems, 1998, 30(1-7): 107-117

[2] Kleinberg J M. Authoritative sources in a hyperlinked environment[J]. Journal of the ACM, 1999, 46(5): 604-632

[3] Zhou Y, Croft W B. Document quality models for web ad hoc retrieval[C]// Proceedings of the ACM Fourteenth Conference on Information and Knowledge Management. 2005: 331-332

[4] Jurczyk P, Agichtein E. Discovering authorities in question-answer communities using link analysis[C]// ACM Conference on Information and Knowledge Management(CIKM). 2007

[5] Jurczyk P, Agichtein E. Hits on question answer portals: an exploration of link analysis for author ranking[C]// SIGIR (posters). ACM, 2007

[6] Zhang M S J. Ackerman Lada Adamic. Expertise Networks in Online Communities: Structure and Algorithms[C]// WWW 2007. Banff, Alberta, Canada, May 2007

[7] Jeon J, Croft B W, Lee J H, et al. A framework to predict the quality of answers with non-textual features[C]// Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, NY, USA: ACM Press, 2006: 228-235

[8] Scott J P. Social Network Analysis: A Handbook[M]. SAGE Publications, January 2000

[9] Zheng Zhi-ping. AnswerBus Question Answering System[C]// Human Language Technology Conference (HLT 2002). San Diego, CA, March 2002: 24-27

[10] Agichtein E, Liu Y. Modeling Information-seeker Satisfaction in Community Question Answering[J]. ACM Transactions on Knowledge Discovery from Data, 2009, 3(2)

[11] Athenikos S J, Han H, Brooks A D. A Framework of a Logic-based Question-answering System for the Medical Domain (LOQAS-Med)[C]// SAC'09. Honolulu, Hawaii, U. S. A. March 2009: 8-12

[12] Tu Xu-dong, Wang Xin-jing, Feng Dan, et al. Ranking Community Answers via Analogical Reasoning[C]// WWW 2009. Madrid, Spain ACM, April 2009: 20-24

[13] Jijkoun V, de Rijke M-T. Retrieving Answers from Frequently Asked Questions Pages on the Web[C]// CIKM. 2005

[14] Lai Y S, Fung K A, Wu C H. Faq *mining via list detection*[C]// Proceedings of the Workshop on Multilingual Summarization and Question Answering. 2002

[15] Cong Gao, Wang Long, Lin C-Y, et al. Finding Question-Answer Pairs from Online Forums[C]// SIGIR 2008. Singapore, July 2008: 20-24

[16] Jeon J, Croft W B, Lee J H. Finding Similar Questions in Large Question and Answer Archives[C]// CIKM'05. Bremen, Germany. 2005

[17] Liu Jing-jing, Cao Yun-bo. Low-quality Product Review Detection in Opinion Summarization[C]// Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Prague, 2007: 334-342

[18] Hoang L, Lee J-T, Song Y-I, et al. A Model for Evaluating the Quality of User-created Documents[C]// Li H, et al., eds. AIRS 2008. LNCS 4993. 2008: 496-501

[19] Rudner L M, Liang T. Automated essay scoring using bayes[J]. Journal of Technology, Learning, and Assessment, 2002, 1(2)

[20] Agichtein E, Castillo C, Donato D, et al. Finding high-quality content in social media[C]// Proceedings of the International Conference on Web Search and Web Data Mining Palo Alto. California, USA, 2008

[21] Shah C, Pomerantz J. Evaluating and Predicting Answer Quality in Community QA[C]// SIGIR'10. Geneva, Switzerland, July 2010: 19-23

[22] 郑实福,刘挺,秦兵,等. 自动问答综[J]. 中文信息学报,2002,16(6)

[23] 张亮,王树梅,黄河燕,等. 面向中文问答系统的问句句法分析[J]. 山东大学学报:理学版,2006,41(3)

[24] 张仰森,曹元大,俞士汶. 基于规则与统计相结合的中文文本自动查错模型与算法[J]. Journal of Chinese Information Processing, 2006, 20(4)

[25] Broder A, Kumar R, Maghoul F, et al. Graph structure in the Web[J]. Computer Networks, 2000: 33(1-6): 309-320