•信息工程•

DOI:10.15961/j.jsuese.201800128

问答社区中基于问题粒度的用户专业性预测方法

朱　敏[1],田　伟[1],彭　第[1],苏亚博[1],牛　颢[2]

(1.四川大学 计算机学院,四川 成都 610065;2.四川省计算机研究院,四川 成都 610041)

摘　要:在线问答社区中大量问题等待回答时间过长、高质量回答数极少,对社区用户在具体问题上的专业程度进行度量具有现实需求。现有的基于链接分析和基于文本分析等方法多集中在社区和话题粒度的专业性度量,并未深入到问题粒度。针对上述问题,定义了问答社区中基于问题粒度的用户专业性概念,在此基础上提出了基于问题粒度的用户专业性预测方法,包括用户专业性度量方法和用户专业性预测模型。该预测方法先利用问答社区中社区用户对回答质量的评价机制,在问题粒度上为用户建立专业性度量;再基于矩阵分解,融合用户偏差、问题偏差以及用户已回答问题集的隐含反馈等信息,构建用户在问题粒度上的专业性预测模型,进而预测用户在待回答问题上的专业程度。利用知乎问答社区互联网话题下的问答数据集,设计了与前述两种主流方法的对比实验。实验结果表明,提出的用户专业性度量方法可以有效地度量用户在具体问题上的专业程度,基于此方法构建的用户专业性预测模型具有更高的预测精度。

关键词:在线问答社区;问题粒度;用户专业性;矩阵分解;预测模型

中图分类号:TP393　　**文献标志码**:A　　**文章编号**:2096-3246(2019)01-0173-08

Method of Predicting User Professionalism Based on Question Granularity in Community Question Answering

ZHU Min[1], *TIAN Wei*[1], *PENG Di*[1], *SU Yabo*[1], *NIU Hao*[2]

(1.College of Computer Sci., Sichuan Univ., Chengdu 610065, China; 2.Sichuan Inst. of Computer Sciences, Chengdu 610041, China)

Abstract: In online community question answering (CQA), many raised questions under go long response time and lack high quality answers. There is a realistic need to measure the community users' professionalism degree on a specific problem. To date, previous methods based on link analysis or text analysis focused on only the professional metrics of community and topic, and did not fully investigate the question granularity. To address this issue, the concept of user professionalism based on question granularity in CQA was defined, and a prediction method for user professionalism based on question granularity was proposed, including a measurement method and a prediction model. Based on the community users' evaluation mechanism of answering qualities, the prediction method established professional metrics of users on the question granularity. Integrating together the user bias, the problem bias and the latent feedback of the question set that users answered, a model on problem granularity based on matrix factorization is constructed to predict how professional the user is in answering questions. By using the question–answer (QA) dataset under topics of Internet in Zhihu, comparative experiments with two mainstream methods were conducted. The results showed that the proposed measurement method of evaluating the degree of user professionalism is effective, and the prediction model has higher prediction accuracy.

Key words: community question answering; question granularity; user professionalism; matrix factorization; prediction model

随着以用户为中心的Web2.0的发展,帮助人们快速准确获取信息的Yahoo! Answers、Quora、知乎、百度知道等在线问答社区应运而生[1−2]。随着社区用户规模增大,问题与回答的数量骤增,尽管有搜索、待解决问题列表等工具可以帮助用户快速查询想要回答的问题,但大部分问题获得解答仍然需要很长时间,回答质量也参差不齐。由于问答社区非常依赖用户参与,问题长期得不到解决将严重影响用户积

收稿日期:2018 − 02 − 04

基金项目:国家自然科学基金资助项目(61572332);四川省重点研发项目资助(2018GZ0171)

作者简介:朱　敏(1971—),女,教授,博士. 研究方向:智能信息处理. E-mail:zhumin@scu.edu.cn

网络出版时间:2019 − 01 − 16 11 : 07 : 04　　**网络出版地址**:http://kns.cnki.net/kcms/detail/51.1773.TB.20190115.1417.002.html

极性，降低整个社区的效率。因此，准确评价用户对于某个问题的专业性，进而为待回答的问题推荐专业的回答者，使问题快速、高质量地被解答，对问答社区的发展至关重要[3]。

目前，在线问答社区的用户专业性评价方法有很多，主要分为基于链接分析的方法[4]和基于文本分析的方法[5]。基于链接分析的方法，通过建模用户之间的问答关系发现专家用户，主要使用置信理论[6]、超链接诱导主题搜索（HITS）算法[7]、网页排名（PageRank）算法[8−9]、共现关系[10−11]、深度学习[12]等对社区关系网络进行链接分析。基于文本分析的方法，通过建模用户问答文档的文本信息发现用户的擅长内容，从而进行专家推荐。计算用户节点的3个值：专业性、响应得分以及引用排名，使用TF-IDF[13]、主题模型[14−17]、语言语义模型[18]等方法计算配置信息与问题间的匹配程度。在商业化的问答社区中，通常会考虑融入更多的因素以实现改进的专家推荐算法。San Pedro等[19]利用排序学习的思想，提出一种基于LDA的RankSLDA模型，应用于对回答者的推荐排序。Zhou等[17]提出了一种最敏感概率模型，通过扩展PageRank算法，将链接和用户分析合并到一个统一的框架。Li等[20]构建Tag-LDA对用户主题分布进行建模，从而挖掘专家用户。Geerthik等[21]根据问答社区中的常用参数提出了一个模型领域专家级排名并在Quora数据上进行验证。

上述研究涉及的用户专业性评判方法主要存在以下问题：基于用户在整个社区的排名度量用户的专业性，而没有考虑用户对于具体领域和具体问题的专业性；仅关注用户兴趣与问题之间的匹配程度，忽视了回答质量与问题之间的关系，从而未能充分利用社区用户对回答质量的评价机制衡量用户在具体问题上的专业性；此外，类似于Yahoo! Answers的商业化问答社区，其目的是最大化用户浏览问题的概率，并不会对用户是否能产生高质量回答进行度量。

针对以上不足，首先，作者将问答社区中的用户专业性划分为3个粒度，即社区粒度、话题粒度和问题粒度。其中，问题粒度，相比于传统的社区粒度（用户在社区中的综合专业性）以及话题粒度（用户在某个领域上的专业性），更适合度量用户在具体问题上的专业程度。然后，提出了一种基于问题粒度的用户专业性预测方法。该方法包括基于社区用户对回答质量评价机制的用户专业性度量方法，以及基于矩阵分解的用户专业性预测模型。在度量阶段，基于对回答质量的评价，考虑问题热度、问题质量等因素对回答的质量评价造成的偏差，对用户在回答上的投票数进行标准化处理，从而构建用户在问题上的专业性度量。在预测阶段，考虑用户偏差等信息，建立基于矩阵分解的用户专业性预测模型，并经过度量阶段结果的训练，从而对用户在未回答问题上的专业性进行预测。

1 基于问题粒度的用户专业性

在线问答社区中，主要存在3类基本实体：用户、问题与回答，对于一个问题，每个用户至多仅能回答一次。用户的专业性主要通过在社区中回答问题的方式表现，但是，对用户专业性的度量则需要区分不同的层次。比如，在某一领域专业的用户，并不一定对其他领域也很专业，同时，社区中综合排名靠前的用户也有其不熟悉的问题。因此，用户专业性的度量粒度由高到低可以分为：用户在社区中的综合专业性，即社区粒度的专业性；用户在某个话题（领域）上的专业性，即话题粒度的专业性；针对具体问题，用户对该问题的专业程度，即问题粒度的专业性。按照此分类方式，基于链接分析的方法是度量用户在社区粒度上的专业性，基于文本分析的方法是度量用户在话题粒度上的专业性。相比以上两者，问题粒度更细，更能反映用户在具体问题上的专业性，作者定义用户在问题上的专业性为：

给定用户集合$U=\{u_1,u_2,\cdots,u_n\}$，问题集合$Q=\{q_1,q_2,\cdots,q_m\}$，则用户u_i在问题q_j的专业性为e_{ij}，其中，$e_{ij}\in[0,1]$，e_{ij}越大表示用户在该问题上越专业，回答质量也就越高。

在问答社区中，基于问题粒度的用户专业性具体表现在该用户回答了什么问题及这些回答得到了怎样的评价。回答问题对于表现专业性是一种很强的信号，然而，用户并不一定回答了所有问题。对于没有回答某个问题的用户，并不能说明这些用户对该问题不擅长，现实中有很多原因造成擅长该问题的用户没有回答，比如，没有机会看到该问题或者该问题已经得到了较好的解答等。由于从问答社区中观测到的回答只是所有〈问题，用户〉回答空间中很小的一部分，所以，需要对用户在问题上的专业性建立模型进行度量，使得对于任意问题$q_j\in Q$，任意用户$u_i\in U$，模型都能够给出用户u_i在问题q_j上的专业性e_{ij}。

基于以上分析，定义用户的专业性建模任务为：

对于用户集合$U=\{u_1,u_2,\cdots,u_n\}$，问题集合$Q=\{q_1,q_2,\cdots,q_m\}$，给定回答数据集$\{a_{ij}|$如果用户u_i回答了问题$q_j\}$：

1）对于已知的〈问题，用户〉对，构建用户在问题上的专业性度量；

2）建立专业性预测模型，对于未知〈问题，用户〉对，预测其专业性。

2　基于问题粒度的用户专业性预测方法

基于问题粒度的用户专业性预测方法的框架如图1所示。该方法首先利用社区中已知的〈问题q_j，用户u_i〉对，构建用户的专业性度量；然后利用度量结果并采用基于随机梯度下降的算法训练用户专业性预测模型。

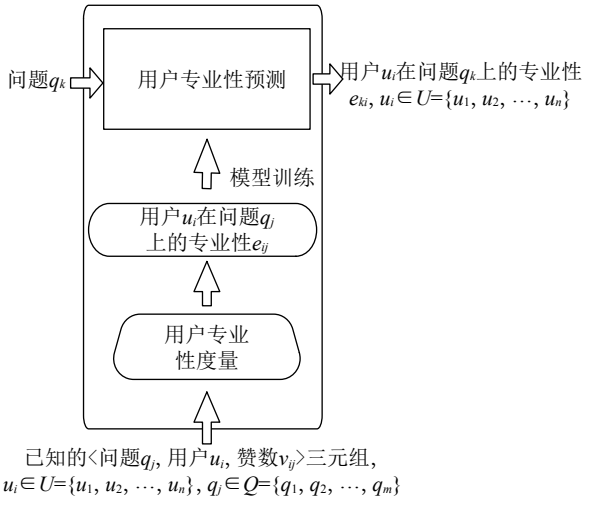

图 1　预测框架

Fig. 1　Frame of prediction

2.1　用户专业性度量方法

问答社区中，社区用户能够对回答进行“点赞”等反馈操作，这种评估机制客观地反映了回答者在具体问题上的专业性。将用户u_i在问题q_j上的回答获得的点赞数（简称“赞数”）记为v_{ij}，$v_{ij} \in [0,+\infty)$，对于用户u_r与用户u_s，如果在问题q_j下的回答获得的投票数$v_{rj} > v_{sj}$，则其专业性也应该满足$e_{rj} > e_{sj}$。因此，定义用户u_i在问题q_j上的专业性e_{ij}为：

$$e_{ij} = f(v_{ij}) \tag{1}$$

式中：$e_{ij} \in [0,1]$；选取满足以下条件的函数为转化函数$f(x)$：

1）定义域为$(-\infty,+\infty)$，值域为$[0,1]$；

2）函数二阶可导，且单调递增。

满足上述两个条件的转化函数有很多，作者使用S型函数（即Sigmoid函数，其图像如图2所示）作为转化函数：

$$f(x) = \frac{1}{1+\mathrm{e}^{-x}} \tag{2}$$

S型函数的函数值随着自变量的增大而趋近于1，且当x趋于$+\infty$时，函数的斜率越来越小。S型函数的这种性质符合赞数的直观理解，即赞数越多，回答的质量越高，用户的专业性越高，而且当赞数达到一定程度后，赞数的微小变化对专业性度量的影响可以忽略，比如，赞数为1 000与赞数为1 010的回答可以认为在专业性上没有差别。

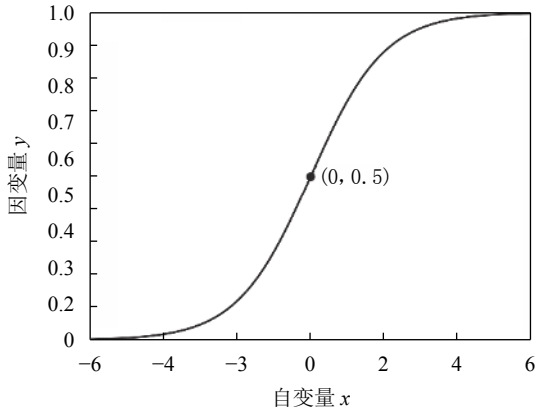

图 2　Sigmoid函数

Fig. 2　Sigmoid function

由图2可知，当x取值大于6时，函数值几乎为1。如果直接使用赞数作为自变量，当赞数大于6后，用户的专业性度量值接近1，同一问题下大部分人的专业性几乎没有差异。但是，当两个用户所获得的赞数差别较大时，专业性应该表现出较大的差异，比如，赞数为10与赞数为100的回答在专业性上是有明显差别的。

假设所有用户回答了所有问题，获得的平均赞数为v_{avg}：

$$v_{\mathrm{avg}} = \frac{1}{|U||Q|}\sum_{u_i \in U}\sum_{q_j \in Q} v_{ij} \tag{3}$$

v_{avg}代表了全体社区用户的平均水平，v_{ij}/v_{avg}则表示用户回答获得的赞数相比于平均水平的倍数。比如，回答的平均赞数为10，赞数为10的回答与平均水平一致，赞数为100的回答是平均水平的10倍。

因此，式（1）更新为：

$$e_{ij} = f\left(\frac{v_{ij}}{v_{\mathrm{avg}}}\right) \tag{4}$$

通常情况下，用户并没有回答社区中的所有问题，从问答社区中观测到的回答可以看作是所有用户在所有问题上全部回答的抽样，因此可使用问答社区中观测到的所有回答的平均赞数近似计算v_{avg}：

$$v_{\mathrm{avg}} = \frac{1}{\sum_{u\in U}\sum_{q\in Q} I(u,q)}\sum_{u_i \in U}\sum_{q_j \in Q \wedge I(u_i,q_j)=1} v_{ij} \tag{5}$$

式中，$I(u,q)$为指示函数，定义为：

$$I(u,q) = \begin{cases} 1, \text{若用户}u\text{已回答问题}q; \\ 0, \text{若用户}u\text{未回答问题}q \end{cases} \tag{6}$$

用户回答所获得赞数的高低不仅与用户的专业

性相关，还与问题的关注度有关。问题的关注度越高，其回答也倾向于获得更高的赞数。比如，在热门问题下一些质量相对不是很高的回答，其赞数会比非热门问题下质量较高的回答获得的赞数更多。因此，针对不同问题，回答所获得赞数的绝对大小不具备可比性，所以，在度量用户的专业性时存在关注度偏差。

对于一个问题q_j，假设所有用户回答该问题获得的平均赞数为v_j：

$$v_j = \frac{1}{|U|}\sum_{u_i \in U} v_{ij} \tag{7}$$

则式(4)更新为：

$$e_{ij} = f\left(\frac{v_{ij} - v_j}{v_{\text{avg}}}\right) \tag{8}$$

v_j代表了所有用户在该问题上的平均表现，即整个社区用户在该问题上所获得赞数的平均水平；$(v_{ij} - v_j)$则修正了问题的关注度对用户专业性度量的偏差。同理，一个问题不会得到社区中所有用户的回答，因此v_j不能直接计算，将从问答社区中观测到的该问题下的全部回答视作所有用户在该问题下产生的回答的抽样，通过该问题已产生的全部回答的平均赞数近似计算v_j：

$$v_j = \frac{1}{\sum\limits_{u \in U} I(u, q_j)} \sum_{u_i \in U \wedge I(u_i, q_j)=1} v_{ij} \tag{9}$$

综上所述，用户u_i在问题q_i上的专业性度量为：

$$e_{ij} = \frac{1}{1 + \mathrm{e}^{-\left(\frac{v_{ij} - v_j}{v_{\text{avg}}}\right)}} \tag{10}$$

式中，v_{avg}由式(5)计算得到，v_j由式(9)计算得到。

2.2 预测模型构建

给定用户集合U与问题集合Q，已知某用户在部分问题上的专业性，预测该用户在其它问题上的专业性。用f维的隐含因子向量分别表示用户$u \in U$与问题$q \in Q$：$\boldsymbol{P}_u \in \mathbb{R}^f$，$\boldsymbol{S}_q \in \mathbb{R}^f$，将用户与问题转化到统一的$f$维联合隐含因子空间，从而用户与问题之间的交互能够用隐含因子向量内积的形式建模。对于问题q，向量$\boldsymbol{S}_q$的每个元素表示了该问题与相应因子之间的相关程度；对于用户u，向量$\boldsymbol{P}_u$的每个元素表示了该用户在相应因子上的专业性。那么，建立用户u在问题q上的专业性预测模型为：

$$e_{uq} = \boldsymbol{P}_u^{\mathrm{T}} \boldsymbol{S}_q \tag{11}$$

一般认为，社区用户的专业性以整个社区用户专业性的平均水平为基准上下波动，而整个社区中所有用户基于问题粒度的专业性平均水平e_{avg}为：

$$e_{\text{avg}} = \frac{1}{\sum\limits_{u \in U}\sum\limits_{q \in Q} I(u, q)} \sum_{u_i \in U} \sum_{I(u_i, q_j)=1} e_{ij} \tag{12}$$

因此，专业性预测模型更新为：

$$e_{uq} = \boldsymbol{P}_u^{\mathrm{T}} \boldsymbol{S}_q + e_{\text{avg}} \tag{13}$$

问答社区中问题的难度不同，其回答所表现出的专业性也不同。比如，一些问题难度较大，大部分用户可能由于专业性不足而无法回答该问题，那么，回答了该问题的用户相比未回答该问题的用户更能表现出专业性。同理，当问题比较简单时，该问题下的回答所表现出的专业性就没有那么准确。也就是说，问题本身会对回答者的专业性度量产生影响，而这种影响并非由于用户与问题的隐含因子向量之间的交互而产生。因此，对于问题q，添加因子$b_q \in \mathbb{R}$，表示问题本身对用户专业性的预测产生的影响。

类似地，用户回答问题的偏好、态度等用户因素也会对用户专业性的预测产生影响。比如，用户对待每个回答的态度很认真，花费大量精力组织回答，从而该用户回答的质量普遍较高；某些用户可能受限于自身的表达能力，即使该用户对此问题很专业，产生的回答的质量也不是很高。因此，对于用户u，在模型中添加因子$b_u \in \mathbb{R}$，表示用户在专业性之外的自身因素对用户专业性的预测产生的影响。

因此，专业性预测模型更新为：

$$e_{uq} = \boldsymbol{P}_u^{\mathrm{T}} \boldsymbol{S}_q + e_{\text{avg}} + b_q + b_u \tag{14}$$

在问答社区中，用户主要通过回答问题表现自己的专业性，而不同的问题之间存在内在联系，用户已回答的问题集合能够隐含地反映出该用户在其他问题上的专业性。因此，对于问题q，添加隐含因子向量$\boldsymbol{X}_q \in \mathbb{R}^f$，表示用户已回答的问题$q$对基于问题粒度的用户专业性的预测产生的隐含反馈。用户已回答的所有问题对基于问题粒度的用户专业性预测的隐含反馈可表示为：

$$\boldsymbol{M}_u = \sum_{q \in Q(u)} \boldsymbol{X}_q \tag{15}$$

式中，$Q(u)$表示用户u已回答问题的集合。但是，每个用户已回答问题的个数不同，将式(15)进行标准化处理：

$$\boldsymbol{M}_u = \frac{1}{|Q(u)|} \sum_{q \in Q(u)} \boldsymbol{X}_q \tag{16}$$

综上所述，基于问题粒度的用户专业性预测模型为：

$$\hat{e}_{uq} = \left(\boldsymbol{P}_u + \frac{1}{|Q(u)|} \sum_{q \in Q(u)} \boldsymbol{X}_q\right)^{\mathrm{T}} \boldsymbol{S}_q + e_{\text{avg}} + b_q + b_u \tag{17}$$

式中，e_{avg}由式（12）计算得到。

2.3 预测模型优化求解方法

预测模型的目的在于尽可能准确地建模用户u在问题q上的专业性，因此，使用模型预测值与真实值之间的差值平方作为损失函数，同时，在损失函数中加入L2正则项，以提高预测模型的泛化能力。对于用户u在问题q上的专业性，定义预测模型的损失函数为：

$$l(e_{uq},\hat{e}_{uq})=\frac{1}{2}(e_{uq}-\hat{e}_{uq})^2+\frac{1}{2}\lambda\left(\|\boldsymbol{P}_u\|^2+\|\boldsymbol{S}_q\|^2+\sum_{q\in Q(u)}\|\boldsymbol{X}_q\|^2+b_q^2+b_u^2\right) \tag{18}$$

式中：e_{uq}为用户u在问题q上的专业性的真实值，由式（10）计算；$\hat{e}_{uq}$为模型对用户u在问题q上的专业性的预测值，由式（17）计算；λ为L2正则项系数。

此模型的优化求解问题是一个无约束最优化问题，而随机梯度下降法是目前解决该类问题比较常见的方法。对于每个训练样本，计算目标函数关于参数的偏导数，在每次迭代中沿目标函数的负梯度方向更新参数：

$$w\leftarrow w-\alpha\frac{\partial l}{\partial w} \tag{19}$$

式中，w为模型的参数，α为每次迭代的学习步长，l为目标函数。

此模型的损失函数$l(e_{uq},\hat{e}_{uq})$连续可导，可使用随机梯度下降优化算法最小化损失函数，模型优化求解方法的输入数据为：隐含因子向量的维度f、L2正则项系数λ、随机梯度下降的迭代学习步长α及最大迭代次数N。具体步骤如下：

1）对于每个用户u，随机初始化f维的隐含因子向量$\boldsymbol{P}_u$及用户因子b_u；对于每个问题q，随机初始化f维的隐含因子向量$\boldsymbol{S}_q$、$\boldsymbol{X}_q$及问题因子b_q。

2）根据式（12）计算整个社区用户专业性的全局平均值e_{avg}以及每个用户回答的问题集合$Q(u)$。

3）遍历训练集中的每个〈问题q，用户u〉回答对，根据式（10）、（17）分别计算用户u在问题q上专业性的真实值e_{uq}和预测值$\hat{e}_{uq}$；计算损失函数$l(e_{uq},\hat{e}_{uq})$分别关于$\boldsymbol{P}_u$、$\boldsymbol{S}_q$、$\boldsymbol{X}_q$、b_u、b_q的偏导数；最后根据式（19）更新预测模型的参数值。

4）重复执行步骤3），直至模型收敛或训练次数达到最大迭代次数。

预测模型训练的时间复杂度与隐含因子向量的维度f、训练数据集中样本的个数n_{Train}以及迭代次数k相关。训练过程的每步迭代均需计算用户与问题之间f维隐含因子向量的内积，因此模型训练的时间复杂度为$O(k\cdot n_{Train}\cdot f)$。同时，由式（17）可知，模型预测的时间复杂度为$O(n_{Test}\cdot f)$，其中，$n_{Test}$表示测试数据集中样本个数。

3 实验对比与分析

3.1 数据准备

实验使用知乎问答社区的“互联网”话题下2016年1月1日到2017年12月31日两年间的问答数据，其中，共有17 714个问题和427 648个回答，涉及188 997名社区用户。

实验数据集中的数据形式如下：

1）问题q由以下属性描述：*qid*、*askerId*、*title*、*detail*、*tag*、*ts*。其中，*qid*表示问题id，*askerId*表示该问题提出者的用户id，*title*表示该问题的标题，*detail*表示该问题的详细内容，*tag*表示该问题所属的话题标签，*ts*表示该问题的提出时间。

2）回答a由以下属性描述：*aid*、*qid*、*responderId*、*content*、*upvote*、*ts*。其中，*aid*表示回答id，*qid*表示该回答所属的问题id，*responderId*表示回答者的用户id，*content*表示该回答的具体内容，*upvote*表示该回答获得的“赞”数，*ts*表示该回答的产生时间。

在实验数据集中，对于每个问题，将其下的回答按照十折交叉验证的方式随机划分为训练集和测试集。

3.2 用户专业性度量方法评估

3.2.1 度量方法评价标准

不同的用户专业性度量方法，在专业性构建方式上存在差异，导致无法通过数值对不同方法所构建的用户专业性进行比较。因此，采用Spearman排名相关系数比较不同的用户专业性度量方法所给出的用户排名与真实的用户排名之间的差异。

Spearman相关系数在区间[−1,1]上，系数值越大，两者相关性越高。对于每个问题q，用户专业性排名的Spearman排名相关系数的计算式为：

$$r_q=1-\frac{6\sum_{u\in U_{Test}(q)}(r_u-\hat{r}_u)^2}{|U_{Test}(q)|\cdot\left(|U_{Test}(q)|^2-1\right)} \tag{20}$$

式中，$U_{Test}(q)$为测试集中回答问题q的用户集合，r_u为用户的真实排名，$\hat{r}_u$为不同的用户专业性度量方法给出的用户排名。

由于测试集中有多个问题，取这些问题的Spearman排名相关系数的平均值作为最终结果：

$$r=\frac{1}{|Q_{Test}|}\sum_{q\in Q_{Test}}r_q \tag{21}$$

式中，Q_{Test}为测试集中的所有问题集合。

3.2.2 对比算法

1）基于链接分析的用户专业性度量方法

将用户视作网络中的节点，在问题的提出者与回答者之间添加一条提出者指向回答者的有向边，构建用户之间的网络拓扑。使用PageRank算法对构建的网络拓扑进行链接分析：

$$pr_u = PageRank(G,u) \tag{22}$$

式中，G为所构建的网络拓扑。对于网络中的每个用户u都对应一个PageRank值pr_u，以pr_u作为用户专业性排名的依据。

2）基于文本分析的用户专业性度量方法

将某问题文本及该问题下的所有回答文本视作关于该问题的文档，将某用户的所有回答文本及其对应的问题文本视作关于该用户的文档，使用LDA算法对这些文档进行主题提取（主题个数设置为200）。从而，每个问题、每个用户都会对应一个主题分布向量，那么，用户u与问题q的主题匹配度t_{uq}：

$$t_{uq} = \sum_{z\in Z} P(z|u)\times P(z|q) \tag{23}$$

式中，Z表示主题集合，$P(z|u)$与$P(z|q)$分别表示用户u与问题q在主题z下的概率。用户与问题的主题越匹配，表明用户越擅长该问题，因此，将t_{uq}作为用户u在问题q上的专业性度量。

3.2.3 结果分析

由于Spearman排名相关系数的计算与用户数（问题回答数）有关，因此，计算用户专业性度量方法分别在不同回答数的问题集合上的平均Spearman排名相关系数，结果如图3所示。图3中，横轴表示回答数大于等于某数值的问题集合，纵轴表示用户专业性度量方法在相应问题集合上的Spearman排名相关系数。

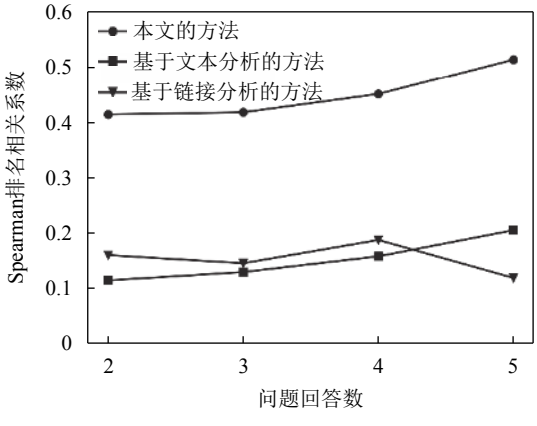

图 3 各度量方法的Spearman相关系数对比

Fig. 3 Comparison of Spearman correlation coefficient of different measurement methods

由图3可以看出：基于链接分析与基于文本分析的用户专业性度量方法性能相近，本文方法在度量基于问题粒度的用户专业性上表现最好；同时，本文方法与基于文本分析的度量方法都具有用户专业性的度量效果随着问题回答数的增多而提升的性质。

3.3 用户专业性预测模型评估

3.3.1 预测模型评价标准

用户专业性预测模型的预测目标是用户在问题上的专业性，模型的预测值与真实值之间的差异越小，模型的性能越好。因此，作者采用均方根误差（root mean square error, *RMSE*）衡量模型的预测精度，*RMSE*数值越小，模型的预测准确性越高。均方根误差的计算式为：

$$RMSE = \sqrt{\frac{1}{|TestSet|}\sum_{(u,q)\in TestSet}(e_{uq}-\hat{e}_{uq})^2} \tag{24}$$

式中，$TestSet$为测试集的样本集合，e_{uq}为用户专业性度量的真实值，$\hat{e}_{uq}$为预测模型的预测值。

3.3.2 对比模型

1）均值模型

均值模型作为基准模型，认为用户在未知问题上的专业性与该用户在已回答问题上的专业性的平均水平相同。对于任意用户u在已回答问题上的专业性的平均水平为：

$$e_{\text{avg}}(u) = \frac{1}{|Q(u)|}\sum_{q\in Q(u)} e_{uq} \tag{25}$$

2）矩阵分解模型

基本的矩阵分解模型预测用户u在问题q上的专业性的方式为：

$$\hat{e}_{uq} = \boldsymbol{P}_u^{\mathrm{T}}\boldsymbol{S}_q \tag{26}$$

其损失函数定义为：

$$l(e_{uq},\hat{e}_{uq}) = \frac{1}{2}(e_{uq}-\hat{e}_{uq})^2 + \frac{1}{2}\lambda\left(\|\boldsymbol{P}_u\|^2 + \|\boldsymbol{S}_q\|^2\right) \tag{27}$$

式中：e_{uq}为用户专业性的真实值；$\hat{e}_{uq}$为模型给出的用户专业性的预测值，由式（26）可得；λ为L2正则项系数。通过随机梯度下降算法训练得到模型参数的最优值。

3）libFM模型

FM（factorization machine）模型是一种矩阵分解的改进模型，使用Rendle等[22]提供的实现工具libFM进行对比实验。

3.3.3 模型参数调优

隐含因子向量的维度f和L2正则项系数λ是影响本文提出的预测模型性能的主要参数。

如图4所示：令$\lambda = 0, f = 10,20,\cdots,100$时，本文的预测模型在训练集上的*RMSE*数值随着f的增大而

减小；当$f>80$时，*RMSE*数值的变化已经趋向平缓。由于模型的时间复杂度会随着f的增大而增高，为平衡模型的准确度和时间复杂度，将本文模型设置为$f=100$。

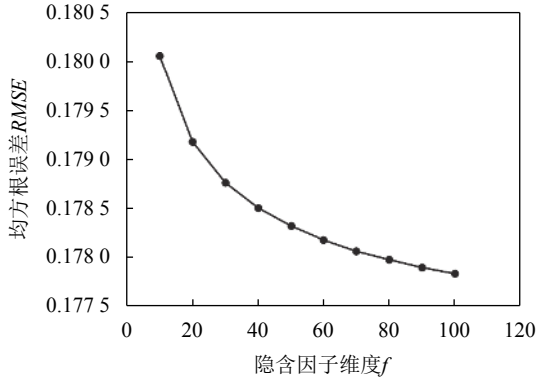

图 4　f取值对模型*RMSE*数值的影响

Fig. 4　Effect of f on *RMSE* of model

如图 5 所示：令$f=100,\lambda=0,0.02,0.04,\cdots,0.40$时，本文的预测模型在训练集上的*RMSE*数值随着λ的增大而先减小后增大，当$\lambda=0.28$时，模型的性能达到最优。

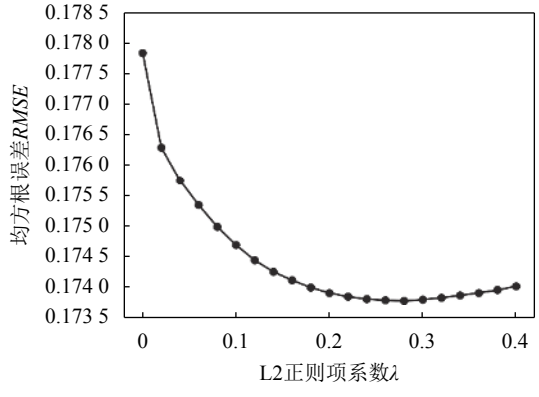

图 5　λ取值对模型*RMSE*数值的影响

Fig. 5　Effect of λ on *RMSE* of model

3.3.4　结果分析

表1为各模型的*RMSE*数值。

表 1　各模型的*RMSE*数值对比

Tab. 1　Comparison of *RMSE* of different models

模型	*RMSE*
均值模型	0.184 3
矩阵分解模型	0.192 9
libFM模型	0.183 1
本文的预测模型	0.173 7

其中，各模型的参数设置如下：

1）均值模型：无参数。

2）矩阵分解模型：隐含因子向量维度$f=20$；L2正则项系数$\lambda=0.01$；随机梯度下降算法的迭代学习步长$\alpha=0.1$，其最大迭代次数$N=20$。

3）libFM模型：运行参数为“libfm -task r -train data/train.dat -test data/test.dat -dim '1,1,30' -out data/result.txt -iter 30”。

4）本文的预测模型：隐含因子向量维度$f=100$；L2正则项系数$\lambda=0.28$；随机梯度下降算法的迭代学习步长$\alpha=0.1$，其最大迭代次数$N=20$。

由表1可以看出，矩阵分解模型的性能低于作为基准的均值模型，libFM模型的性能比均值模型稍好，本文的预测模型效果最好。

4　结　论

作者首先定义了用户在问题粒度上的专业性概念；然后，设计了一种基于问题粒度的用户专业性预测方法，包括用户专业性度量方法和预测模型。在构建用户专业性度量阶段，在社区用户对回答质量评价的基础上，考虑了问题热度、问题质量等因素对回答质量评价造成的影响；在用户专业性预测阶段，基于矩阵分解模型，融合用户偏差、问题偏差以及用户在回答问题上的隐含反馈等信息，建立了专业性预测模型。由实验结果可知，本文方法可以较准确地度量并预测用户在某个问题上的专业性，进而将问题推荐给合适的专家用户，使该问题得到较高水平的解答。下一步研究中，考虑将问答社区用户之间关注与被关注的社交关系信息融合到用户专业性预测方法中，以进一步提高预测结果的准确性。

参考文献：

[1] Adamic L A,Zhang Jun,Bakshy E,et al.Knowledge sharing and Yahoo answers:Everyone knows something[C]//Proceedings of the 17th International Conference on World Wide Web.New York:ACM,2008:665–674.

[2] Mao Xianling,Li Xiaoming.A survey on question and answering systems[J].Journal of Frontiers of Computer Science and Technology,2012,6(3):193–207.[毛先领,李晓明.问答系统研究综述[J].计算机科学与探索,2012,6(3):193–207.]

[3] Dror G,Maarek Y,Szpektor I.Will my question be answered?Predicting “question answerability” in community question-answering sites[C]//Proceedings of the 2013th Joint European Conference on Machine Learning and Knowledge Discovery in Databases.Heidelberg:Springer-Verlag,2013:499–514.

[4] Neshati M,Fallahnejad Z,Beigy H.On dynamicity of expert finding in community question answering[J].Information Processing and Management,2017,53(5):1026–1042.

[5] Rafiei M,Kardan A A.A novel method for expert finding in online communities based on concept map and PageRank[J].Human-centric Computing and Information Sciences,

2015,5:10.

[6] Attiaoui D,Martin A,Yaghlane B B.Belief measure of expertise for experts detection in question answering communities:Case study stack overflow[J].Procedia Computer Science,2017,112:622–631.

[7] Jurczyk P,Agichtein E.Discovering authorities in question answer communities by using link analysis[C]//Proceedings of the Sixteenth ACM Conference on Information and knowledge Management.New York:ACM,2007:919–922.

[8] Zhang Jun,Ackerman M S,Adamic L.Expertise networks in online communities:Structure and algorithms[C]//Proceedings of the 16th International Conference on World Wide Web.New York:ACM,2007:221–230.

[9] Zhou Guangyou,Lai Siwei,Liu Kang,et al.Topic-sensitive probabilistic model for expert finding in question answer communities[C]//Proceedings of the 21st ACM International Conference on Information and Knowledge Management.New York:ACM,2012:1662–1666.

[10] Zhu Hengshu,Chen Enhong,Xiong Hui,et al.Ranking user authority with relevant knowledge categories for expert finding[J].World Wide Web,2014,17(5):1081–1107.

[11] Liu Hanqing,Zhu Min,Su Yabo,et al.A collaborative prediction model for user interest shift feature[J].Journal of Sichuan University(Natural Science Edition),2016,53(3):548–554.[刘汉清,朱敏,苏亚博,等.一种考虑用户兴趣转移特征的协同预测模型[J].四川大学学报(自然科学版),2016,53(3):548–554.]

[12] Zhao Zhou,Yang Qifan,Cai Deng,et al.Expert finding for community-based question answering via ranking metric network learning[C]//Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence.New York:AAAI Press,2016:3000–3006.

[13] Davitz J,Yu Jiye,Basu S,et al.iLink:Search and routing in social networks[C]//Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.New York:ACM,2007:931–940.

[14] Liu Xuebo,Ye Shuang,Li Xin,et al.ZhihuRank:A topic-sensitive expert finding algorithm in community question answering websites[M]//Advances in Web-Based Learning-ICWL 2015.Cham:Springer,2015:165–173.

[15] Riahi F,Zolaktaf Z,Shafiei M,et al.Finding expert users in community question answering[C]//Proceedings of the 21st International Conference on World Wide Web.New York:ACM,2012:791–798.

[16] Sahu T P,Nagwani N K,Verma S.Multivariate Beta mixture model for automatic identification of topical authoritative users in community question answering sites[J].IEEE Access,2016,4:5343–5355.

[17] Zhou Guangyou,Zhao Jun,He Tingting,et al.An empirical study of topic-sensitive probabilistic model for expert finding in question answer communities[J].Knowledge-Based Systems,2014,66:136–145.

[18] Zhou Guangyou,Zhou Yin,He Tingting,et al.Learning semantic representation with neural networks for community question answering retrieval[J].Knowledge-Based Systems,2016,93:75–83.

[19] San Pedro J,Karatzoglou A.Question recommendation for collaborative question answering systems with RankSLDA[C]//Proceedings of the 8th ACM Conference on Recommender Systems.New York:ACM,2014:193–200.

[20] Li Hai,Jin Songchang,Li Shudong.A hybrid model for experts finding in community question answering[C]//Proceedings of the 2015 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery.Xi'an:IEEE,2015:176–185.

[21] Geerthik S,Gandhi K R,Venkatraman S.Domain expert ranking for finding domain authoritative users on community question answering sites[C]//Proceedings of the 2016 IEEE International Conference on Computational Intelligence and Computing Research.Chennai:IEEE,2016:1–5.

[22] Rendle S.Factorization machines with libFM[J].ACM Transactions on Intelligent Systems and Technology,2012,3(3):Article No.57.

（编辑　赵　婧）

引用格式: Zhu Min,Tian Wei,Peng Di,et al.Method of predicting user professionalism based on question granularity in community question answering[J].Advanced Engineering Sciences,2019,51(1):173–180.[朱敏,田伟,彭第,等.问答社区中基于问题粒度的用户专业性预测方法[J].工程科学与技术,2019,51(1):173–180.]