计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200500024

基于用户关系的在线问答平台用户重要性评估方法

李 霄 曲 阳 李 辉 郭世凯
大连海事大学信息科学技术学院 辽宁 大连 116026
(lixiao2048@163.com)

摘 要 在线问答平台日益成为互联网用户知识获取的重要途径,随着其用户数量的迅速增长,平台中重要用户的识别难度逐渐增大,用户提出的大量问题也得不到有效回答,严重影响了在线问答平台的用户体验。针对上述问题,将在线问答平台中用户提出问题和回答问题看作一种社交网络行为,并且根据这些行为构建用户关系网络,在此基础上提出了一种基于用户关系网络的用户重要性评估方法,用来识别平台中的重要用户。通过对 Stack Overflow 数据集的实验分析表明,该方法得出的用户重要性排名与在线问答平台中的实际情况相符合,生成排名结果相对稳定,通过用户重要性排序结果对问题进行推荐可以提高问题回答效率。应用该用户重要性评估方法,设计和开发了一个在线问答平台,案例分析表明该方法能够识别出在线问答平台中的重要用户,可以增强用户知识获取的体验。

关键词:在线问答平台;社交网络;用户重要性排序;问题推荐
中图法分类号 TP393

User Importance Evaluation for Q&A Platform Based on User Relations

LI Xiao,QU Yang,LI Hui and GUO Shi-kai
Information Science and Technology College,Dalian Maritime University,Dalian,Liaoning 116026,China

Abstract Q&A has increasingly become an important platform of acquiring knowledge for WWW users. As the number of the users rapidly increases,the identification of the important users becomes more and more difficult,and more and more questions cannot be answered in Q&A platforms. Thus it seriously affects the user experience. Aiming to solve this problem,we regard the questions and answers of users in the Q&A platform as a kind of social network behavior,and build a user relationship network based on these behaviors. On this basis,we present an evaluation of user importance ranking based on the user relationship network,and further identify the important users of the platforms. Experimental studies based on data set of Stack Overflow show that,the results produced by the user important ranking is consistent with the actual ranking lists,and the produced ranking results are relatively stable. Furthermore,the ranking results can be used for improving the question recommendation. Applying the user importance ranking measurement,we designed and developed a Q&A platform. Empirical studies show that this ranking method can identify the important users from Q&A platform,and improve the user experience of knowledge acquirement.

Keywords Q&A platform,Social network,User importance ranking,Question recommendation

1 引言

目前,在线问答平台已经成为互联网用户获取知识的重要媒介,用户比以往更愿意主动去搜索自己感兴趣的信息,并在感兴趣的领域中提出问题,期待得到令自己满意的回答[1]。在线问答平台中,用户通过互联网媒介组成一个互帮互助的社交协作网络,用户既可以作为提问者,也可以作为回答者,用户关系更加密切,在线问答平台上体现出了较好的用户体验。

在线问答平台的兴起和发展大致经历了两个阶段。前一阶段中用户通过搜索解决自己想了解的问题,代表性平台有 Google Answers 等。后一阶段是用户增加了对相关领域知识的提问和问答,代表性平台有 Stack Overflow、知乎等[2]。以面向程序员用户的在线问答平台 Stack Overflow 为例,因其丰富的功能和良好的用户体验,该平台自 2008 年上线之后迅速在程序员群体中普及,活跃用户众多。据统计,在 2009 年 2 月 18 日至 2009 年 6 月 7 日期间,共有 263541 条关于问题的回答,提出问题和回答问题的用户共有 43509 人。

随着在线问答平台用户数量的剧增,对重要用户的识别和关注越来越受到在线问答平台的重视,因为重要用户在平台中的活跃程度较高,尤其是在目前存在大量问题得不到及时回复的情况下[3],重要用户的积极回答体现出了不可替代的作用。在线问答平台中,问题的解决都依赖于用户的回答,因此作为回答者的用户在平台上发挥了至关重要的作用[4]。例如,一个活跃的用户对于问题解决的贡献程度可能就大于

基金项目:国家自然科学基金(61602077,61902050);中国博士后科学基金(2020M670736);中央高校基本科研业务费项目(3132019355);下一代互联网技术创新项目(NGII20181205,NGII20190627)
This work was supported by the National Natural Science Foundation of China (61602077,61902050),Fellowship of China Postdoctoral Science Foundation(2020M670736),Fundamental Research Funds for the Central Universities (3132019355) and Next-Generation Internet Innovation Project of CERNET (NGII20181205,NGII20190627).
通信作者:李辉(li_hui@dlmu.edu.cn)

一个沉寂的用户，一个在平台中活跃多年的重要用户所做的贡献也应该远远大于一个普通的用户[5]。因此，能够准确度量用户的重要性，并识别出重要用户是非常重要的。

评估在线问答平台中用户的重要性，最简单直接的方法就是统计每个用户的提问数、回答数，或者用户的评分。在统计数据基础上，采用预测和推荐算法，如协同过滤算法[6]，来对用户的重要性排序。因数据规模往往较大，也会采用某些并行计算方法，如 MapReduce，来对用户重要性进行计算和排序[7]。但是，如果涉及专业领域知识的问题，就需要具备该领域知识的用户回答该问题，对这些用户的重要性进行评估时，还需要考虑对用户领域贡献的行为进行重要性评价[8]，或者采用文本挖掘的方法[9]。

采用这种方法能够识别出大部分的重要用户，但是会遗漏一些统计数据排名不靠前，但是回答问题质量很高的用户。这些用户往往难以用统计数据来体现其重要性，但是对很多问题的解答起着重要作用。从社交网络的角度来看，这类用户所代表的节点往往度值不高，但是其邻居节点却往往也都是重要节点。因此，采用社交网络建模方法来制定一系列用户重要性识别规则成为了一种可行方法[10-11]。另外，采用 PageRank 等节点重要性排序方法也能够取得较合理的排序结果[12]。但是，这类方法对用户间的问题回答频度未做全面考虑，且评估结果的成败也部分取决于模型参数设置的大小，在线问答平台实际应用中的效果并不理想[13]。

基于以上原因，在采用用户关系的社交网络模型基础上，考虑用户间问题回答的频度，综合用户邻居的重要性对其的影响，设计一种在线问答平台用户重要性评估方法，评估用户对平台的贡献，并识别出平台中的重要用户。为了验证该方法对重要用户的识别效果，我们应用该方法开发了一个在线问答平台，通过案例分析验证了该方法的有效性。如果应用该方法作为问题推荐的依据，将用户提出的问题推送给重要用户，预期能加快问题的回答效率，进而提升在线问答平台的用户体验。

2　评估方法描述

根据在线问答平台上用户的提问与回答信息构建用户关系网络，在此基础上设计用户重要性评估方法来量化评估用户的重要性，对用户进行重要性排序，以评估用户在在线问答平台的贡献。

2.1　在线问答平台用户关系网络

在进行用户重要性评估计算之前，需要将用户问题中的相关信息提取出来并建立用户关系网络。在用户关系网络中，用户的角色有两个，分别是问题的提交者和回答者。首先，提取每个问题中提交者与回答者的信息。以 Stack Overflow 平台上 Que_id 1 和 Que_id 2 两个问题为例，提取的信息如表 1 所列。

表 1　问题中提问者与回答者的关系

Table 1　Relationship between submitters and repliers in questions

User	Que_id:1	Que_id:2
Questioner	Danica	Danica
Commentator 1	Danica	Hagley
Commentator 2	Jacklyn	Vanessa
Commentator 3	June	Jean
Commentator 4	Hagley	Danica
Commentator 5	Jean	Jacklyn

通过表 1 可以看到，Que_id 为 1 和 Que_id 为 2 的两个问题分别由两个提问者提问，同时每个问题又分别被回答了 5 次。在这两个问题中，出现了相同的回答者，即 Jean 和 Jacklyn 等。Que_id 为 1 的回答者有 5 位，分别为 Danica，Jacklyn，June，Hagley，Jean；Que_id 为 2 的回答者有 5 位，分别为 Hagley，Vanessa，Jean，Danica，Jacklyn。

当某一位用户回答了另一位用户提交的问题或者回答后，可以认为这两个用户之间产生了一条有向连接。图 1 是以 Que_id 1 和 Que_id 2 中的提问者和回答者之间的关系为例建立的用户关系网络。例如，Que_id 1 提问的回答中，根据表 1 可知 Jacklyn 回答了 Danica，因此在图 1 中存在一条有向边，方向从 Jacklyn 指向 Danica，June 继续回答了 Jacklyn 的回答，因此图 1 中便会增加一条从 June 指向 Jacklyn 的边。

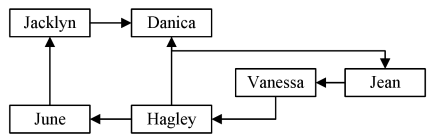

图 1　依据 Que_id 1 和 Que_id 2 问题构建的用户关系网络

Fig. 1　User relationship network produced by Que_id 1 and Que_id 2

因此，建立用户关系网络的流程如下：1）将在线问答平台中参与提问或问答的每一位用户抽象为节点；2）根据每位用户回答问题的行为，建立一条回答者指向被回答者的有向连接（忽略用户自己回答自己的情况）。

在社交网络节点影响力的评估中，一个节点的影响力不仅取决于其邻居节点的数量，也取决于每个邻居节点的影响力。因此，在在线问答平台中，当一个用户提出问题，如果某一个用户回答了该问题，那么该问题的提问者在该用户关系网络中的重要性也会受到该回答者的影响。另一方面，如果一个回答者回答了多个问题，那么该回答者的重要性会平均作用于他（她）回答的这些问题的提问者。

2.2　用户重要性评估方法

目前，许多大型的在线问答平台都有对用户的评分，用来表明用户在该平台的等级，以及以此激励用户继续解决问题。因此，我们用在线问答平台上用户的重要性表示他们对在线问答平台的贡献度。因为在线问答平台上用户众多，存在一些回答问题很少但是回答质量很好的用户，所以对于用户的贡献度的评定不能依靠简单的统计回答数等指标来衡量，于是我们基于用户关系网络设计用户重要性评估方法来量化用户在在线问答平台上的重要性。在线问答平台重要性评估方法的伪代码如算法 1 所示。

算法 1　在线问答平台用户重要性评估方法

输入：在线问答平台用户 i，用户 i 与其他用户之间的连接

输出：在线问答平台用户 i 的重要性 S_i

Step 1　添加一个根节点 g，并与其他用户建立连接构成连通图；

Step 2　定义用户之间的连接权重 W_{ij}，如式（1）所示；

Step 3　设置所有用户初始重要性 $S_i(0)=1$；

Step 4　for t=1 to t_c do //t_c 是收敛次数；使用式（2）计算用户 i 在 t 次时得到的重要性 $S_i(t)$。

在在线问答平台中，用户问题的评论数是用户在此平台的重要性的一个重要依据。如果用户 i 回答了用户 j，那么用户 i 所代表的节点到用户 j 所代表的节点有一条有向边。假定原先是由 n 个节点、m 条边构成的图，为了赋予每个节点表示其重要性的初始值，我们引入一个根节点 g，且节点 g 的重要性为 1。然后创建一条从根节点指向其他节点的边和一条

其他节点指向根节点的边，构成一个强连通图，那么原来的连通图便由 $n+1$ 个节点和 $m+2n$ 条边构成。因此，节点 i 到节点 j 的权重 W_{ij} 定义为式(1)：

$$W_{ij}=\begin{cases}0, & a_{ij}=0 \text{ 且 } i,j\neq g\\ 1, & a_{ij}=1 \text{ 且 } i,j\neq g\\ 1, & j=g\\ k_j^{in} & i=g\end{cases} \tag{1}$$

其中，a_{ij} 表示这个连通图中节点 i 到节点 j 是否存在路径，如果从节点 i 到节点 j 有边，则 $a_{ij}=1$，如果 $a_{ij}=1$，那么 $W_{ij}=1$；如果从节点 i 到节点 j 无边，则 $a_{ij}=0$，如果 $a_{ij}=0$，那么 $W_{ij}=0$；由于我们引入了根节点 g，因此设置 W_{gi} 等于节点 i 的入度 k_i^{in}。另外，从节点 i 指向根节点 g 的重要性 W_{ig} 恒等于 1。

通过构建用户关系网络以及给网络中边和节点赋值，每个节点的初始值是 1，则用户关系网络中用户重要性可以定义为：

$$S_i(t+1)=\sum_{j=1}^{N+1}\frac{W_{ji}}{\sum_{i=1}^{N+1}W_{ji}}S_j(t) \tag{2}$$

其中，$S_i(t)$ 表示节点 i 在网络中迭代 t 次后的重要性，经过计算节点的 $S_i(t)$ 的值越大，那么用户的重要性就越大。该指标计算所需的时间复杂度与用户数量的平方成正比。一般情况下，迭代 30 次便可以得到稳定的重要性，最终每个节点的值就代表用户的重要性，对用户的分值进行排序就可得到用户的重要性排名。

3 实验分析

3.1 实验数据

在采用用户重要性评估方法得到在线问答平台用户重要性，然后对用户重要性进行排序之后，我们将通过实验验证排序结果是否合理。实验采用 Stack Overflow 平台上从 2009 年 2 月 18 日至 2009 年 6 月 7 日期间的数据集。该数据集中的字段属性如表 2 所列。

表 2 问题标签属性

Table 2 Question label attributes

Name	Description
Qid	Unique question id
I	User id of questioner
Qs	Score of the question
Qt	Time of the question (in epoch time)
Tags	A comma-separated list of the tags associated with the question.
Qvc	Number of views of this question (at the time of the datadump)
Qac	Number of answers for this question (at the time of the datadump)
Aid	Unique answer id
J	User id of answerer
As	Score of the answer
At	Time of the answer

在数据集中，*Qid* 代表用户提出问题的唯一标识；*I* 代表问题的提问者的唯一标识；*Qs* 代表问题的得分；*Qt* 代表问题的提出时间；Tags 代表标签，有“html”“R”“mysql”“python”等，每个问题通常使用 2～6 个标签；Qvc 代表问题到目前为止的访问量；Qac 代表问题到目前为止的回答数；*Aid* 代表回答的唯一标识；*J* 代表回答者的唯一标识；*As* 代表答案的得分；*At* 代表答案的回答时间。该数据集一共有 40 359 个用户，其中大多数用户回答的问题少于 50 个。这些用户提出了 83 017 个问题，并且产生了 263 541 个对上述问题的用户回答。

3.2 节点重要性对连通性的影响

根据用户重要性排序结果，我们选取不同数目的较大重要性的用户，验证其重要用户所代表的节点对用户关系网络连通性的影响。通过对比分别使用随机删除和按排名顺序删除这两种对应用户的删除方法，来判断用户重要性对图的连通性的影响，从而确定高重要性的用户在在线问答平台用户中是影响力高的群体。

本文设计了两个不同数量级上的实验：第一个实验中本文删除用户的数目分别为 10，20，30，40，50，60，70，80，90，100，然后通过对比随机删除和按排名顺序删除的两种删除方式产生的子图数量和最大子图用户数，来判断本文方法是否能够找到高重要性的用户；第二个实验中扩大删除的数目，删除用户的数目分别为 200，400，600，800，1 000，1 200，1 400，1 600，1 800，2 000，然后通过对比随机删除和按排名顺序删除的两种删除方式产生的子图数量和最大子图用户数，来判断本文是否找到高重要性的用户。根据实验结果对研究内容中本文提出的改进方法的合理性进行了验证。

本实验的评价标准有 2 个。1)子图数量。在节点删除之前，用户关系网络是一个连通图，删除若干用户之后便可能会导致原图破碎为多个子图。最终，子图数量越多说明图的破碎程度越大，所删除节点在网络中的重要性越高；子图数量越少说明图的破碎程度越小，所删除节点在网络中的重要性越低。2)最大连通子图中节点的个数。一般而言，最大连通子图越小，图的破碎程度越大，说明删除节点在网络中的重要性越高，所代表的用户的重要性越大；最大连通子图越大，所删除节点在网络中的重要性越低，所代表的用户的重要性越小。

图 2－图 5 中的 x 轴是删除的用户数目，图 2 和图 4 的 y 轴是当删除对应数目的用户后产生的子图数，图 3 和图 5 是当删除对应数目的用户后产生的最大连通子图的用户数。图 2－图 3 给出了本文分别删除 10，20，30，40，50，60，70，80，90，100 个用户后，比较随机删除和按排名顺序删除之后产生的连通子图数和最大连通子图的用户数。从图 2 可以看出，随机删除用户数的子图数增长极其缓慢，最终都没有超过 100 个，但是按排名顺序删除的子图数增长就很迅速，当删除到 100 个用户时已经超过了 2 500 个子图数。从图 3 可以看出，随机删除的最大连通子图的用户数减少缓慢，但是按排名顺序删除的最大连通子图的用户数减小的速度很快，最后达到了 3 700 个用户。图 4、图 5 给出了本文分别删除的用户数目为 200，400，600，800，1 000，1 200，1 400，1 600，1 800，2 000 时，比较随机删除和按排名顺序删除之后产生的连通子图数和最大连通子图的用户数。

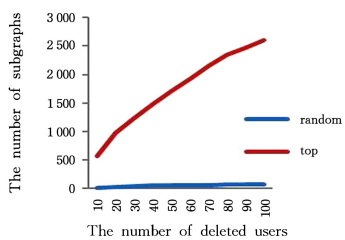

图 2 子图数量与删除用户的关系

Fig. 2 Relationship between the number of subgraphs and deleted users

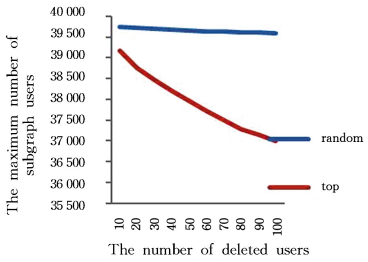

图 3　最大子图用户数与删除用户数的关系

Fig. 3　Relationship between the maximum number of subgraph users and the number of deleted users

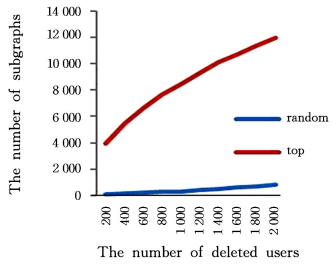

图 4　子图数量与删除用户的关系

Fig. 4　Relationship between the number of subgraphs and deleted users

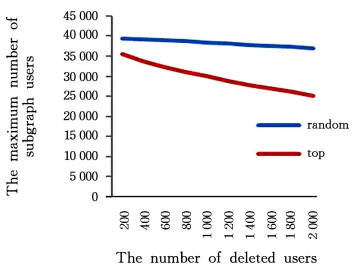

图 5　最大子图用户数与删除用户数的关系

Fig. 5　Relationship between maximum number of subgraph users and the number of deleted users

3.3　用户重要性和问题难易程度的关系

通过分析回答最困难问题的用户排名和用户的最佳答案次数排名，以及计算所得的用户重要性排名，以确定是否找到高重要性的用户。

在实验数据中，每个问题都会有不同的回答者，回答者的回答得分也会有一个最高得分，本实验把最高得分作为本问题的可接受的答案，统计出每个用户的可接受答案的次数作为用户的最佳答案次数，并且对此进行排名；同时为了获得最困难问题的用户排名，引入 *PD* 系数[14]，用 *View* 代表这个问题的浏览数，*Answer* 代表这个问题的回答数，对于每个问题，用其 *View* 除以 *Answer* 代表问题的难度系数。一般情况下，浏览数越多，回答数越少，问题也越难，那么两者的比值也越大，即 *PD* 的分值也越大。*PD* 定义为式(3)：

$$PD=\frac{View}{Answer} \tag{3}$$

用每个用户的 *PD* 值的最高分作为用户回答的最难问题，表示问题的难易程度，然后用 *PD* 值把回答者回答的最难问题进行一个排名，从而得到最难问题的排名。一般而言，用户的重要性高则会偏向于较难回答的问题。重要性高的用户的回答问题相对低重要性用户的回答问题要相对积极一些，因此会更多地参与较难的问题的回答，而且回答的最佳答案的次数也会相对较多。因此，本实验要验证的是用户重要性排名与回答者回答最难问题的排名的相关性，以及用户重要性排名与回答者回答问题的最佳次数排名的相关性。

本实验的评价标准有 2 个：1)用户回答最难问题的排名，即用户的 *PD* 排名；2)用户回答问题的最佳答案次数的排名。分别求出两个评价标准和用户重要性排名的皮尔逊相关性系数，然后判断本文的用户的重要性是否合理。

皮尔逊相关系数[15]用来度量两个变量之间的相关性，皮尔逊系数的取值区间为－1～1。皮尔逊相关性系数为负，代表两个变量之间呈现负相关关系；皮尔逊系数为正，代表两个变量之间呈现正相关关系；皮尔逊相关性系数为 0，则代表两个变量之间无关。皮尔逊相关性系数的计算公式为：

$$r=\frac{\sum_{i=1}^{n}(a_i-\bar{a})(b_i-\bar{b})}{\sqrt{\sum_{i=1}^{n}(a_i-\bar{a})}\sqrt{\sum_{i=1}^{n}(b_i-\bar{b})}} \tag{4}$$

其中，$\bar{a}$ 代表样本 a 的均值。皮尔逊相关性系数是衡量两个变量之间相关性的重要系数。

表 3 列出了用户重要性排名前 20 位用户的 id，以及他们获得最佳答案的数目和所有回答问题所得 *PD* 总分。本文可以看到用户重要性排名越高，*PD* 总分越高，最佳答案的数量越多。例如，id 为 22 656 的用户，其重要性高居榜首，其最佳答案的数量也最多，*PD* 得分也最高。如表 4 所列，用户重要性排名和用户 *PD* 总分排名之间的皮尔逊相关性系数为 0.381 6，用户重要性排名与用户最佳答案个数排名之间的皮尔逊相关性系数为 0.829 3。可以看到，用户重要性排名与用户最佳回答次数排名以及用户 *PD* 总分排名均呈正相关性，且相关性较大，因此用户重要性排名较高的用户在回答问题时表现优秀。

表 3　排名前 20 的用户的最佳回答数和 *PD* 分数

Table 3　Top 20 best answer users' best answers and PD score

User id	The number of best answers	*PD*
22 656	1 095	780
23 354	835	386
69 307	277	492
23 283	485	325
12 950	405	641
76 337	310	290
66 692	324	329
33 708	369	670
60 711	242	323
34 211	277	407
35 092	259	178
18 393	305	408
14 860	306	449
3 043	256	776
10 661	225	221
65 358	253	376
16 076	272	333
28 169	190	244
66 372	175	428
95 810	247	130

表 4　用户重要性排名与两个标准的皮尔逊相关性系数

Table 4　Pearson correlation coefficient between user importance ranking and two standards

Pearson correlation coefficient	The rank of user's *PD*	The rank of user's best answers
user importance ranking	0.381 6	0.829 3

3.4　用户重要性和问题回答指标的关系

通过分析用户重要性和回答的相关指标，来确定是否找

到高重要性用户。本实验包括3个实验：分析用户的回答数排名与用户重要性排名的相关性；分析用户的回答得分排名与用户重要性排名的相关性；分析用户的回答和提问的总得分排名与用户重要性排名的相关性。

在本实验数据中，as 代表回答的得分，qs 代表问题的得分。首先统计用户的回答次数排名，本文在统计用户的回答得分排名时，用户的每一个回答后面都会有一个 as 代表用户回答的答案得分，分别对每个用户的所有回答累计得分总数。统计出用户的回答和提问的总得分，提问的得分就是 qs，然后和上一个统计出来的用户的回答得分相加以确定最终得分，最后列出排名。分别计算用户的回答数排名、用户的回答得分排名、用户的提问回答总得分与用户重要性排名的皮尔逊相关性系数，然后确定三者得分排名与重要性排名的相关性。

本实验的评价标准有3个。1)用户的回答数排名与用户重要性排名的皮尔逊相关性系数。一般而言，用户的重要性越高，那么用户的回答数排名应该越靠前，因此用户的回答数排名与实验算法得分的皮尔逊相关性系数越大，说明用户重要性排序结果越合理。2)用户的回答得分排名与用户重要性排名的皮尔逊相关性系数。同样，用户的重要性越高，越具有影响力，那么他的回答得分的排名应该越靠前，因此用户的回答得分排名与用户重要性排名的皮尔逊相关性系数越大，说明用户重要性排序结果越合理。3)用户的回答提问总得分排名与用户重要性排名的皮尔逊相关性系数。同理，用户的重要性越高，越具有影响力，那么他的回答提问总得分的排名应该越靠前，因此用户的回答提问总得分与用户重要性排名的皮尔逊相关性系数越大，说明用户重要性排序结果越合理。

根据表5的结果还可以看出，用户的回答次数排名和用户重要性排名的皮尔逊相关性系数是0.9489，用户的回答得分排名和用户重要性排名的皮尔逊相关性系数是0.8461，用户的问题和回答的总得分排名与用户重要性排名的皮尔逊相关性系数是0.8801，也就是说回答次数排名、回答得分排名、回答和问题的总得分排名与用户重要性排序结果的排名是正相关的，因此可以说用户重要性排序结果是合理的。

表5 用户重要性与3个标准的皮尔逊相关性系数

Table 5 Pearson correlation coefficient between user importance ranking and three standards

Pearson correlation coefficient	The rank of user's answers	The rank of user's answers score	The rank of user's Q&A score
User importance ranking	0.9489	0.8461	0.8801

表6列出了用户重要性排序前20的用户的算法得分、回答次数、回答得分、回答和问题总得分。在大多数情况下，可以看出用户的重要性高，那么他的回答次数、回答得分以及回答和问题的总得分都会高，如 id 为22656的用户，他的重要性排第一，同样他的回答次数、回答得分、回答和问题的总得分都排名第一。

表6 用户重要性排名前20的用户的回答数和回答分数以及回答和问题的总分数

Table 6 Number of answers and answer scores of the top 20 users and total scores of answers and questions

User id	The score of user importance	The number of answer	Answer scores	The totalscores of Q&A
22656	119.33	1683	7843	7848
23354	99.658	1457	3856	3864
69307	87.247	1033	3184	3217
23283	87.018	1024	3818	3818
12950	81.154	926	1866	1870
76337	80.998	967	962	962
66692	79.286	935	1889	1890
33708	66.767	716	2649	2656
60711	66.415	669	1529	1545
34211	65.813	663	1789	1801
35092	63.41	630	1446	1446
18393	63.115	604	2295	2307
14860	62.856	688	2184	2190
3043	62.202	708	2022	2073
10661	57.107	624	1593	1594
65358	54.303	602	1528	1544
16076	48.178	541	1143	1143
28169	46.475	474	1179	1185
66372	46.3	479	570	570
95810	46.295	575	1184	1190

3.5 用户重要性和回答时间长度指标的关系

本实验验证高重要性用户回答的最佳答案是否是在较短时间内完成。将实验中的83017个问题根据其得到最高分问题答案的时间长短进行排序，然后将其从长到短排序，再把这83017个问题等量划分为10段，第一段就是时间跨度最大的8302个问题，然后以此类推。分别提取根据用户重要性排名前10%，20%，30%，40%，50%的用户，计算在问题的10个分段中，用户回答的问题数量和这些问题数量占分段问题数的百分比。

评价标准为每一个时间跨度分段内用户回答的问题数以及用户回答的问题数占问题总数的百分比。很明显，回答问题数越大，问题所占分段百分比就越大，说明用户在这个分段回答的问题所占比重大。实验结果如表7所列。

表7 最佳回答时间长度与重要性用户的关系

Table 7 Relationship between best answer time and importance user

Time span	Users 10% (count)	Users 10% (radio)	Users 20% (count)	Users 20% (radio)	Users 30% (count)	Users 30% (radio)	Users 40% (count)	Users 40% (radio)	Users 50% (count)	Users 50% (radio)
Part 1	3368	0.4057	4957	0.5971	5825	0.7016	6415	0.7727	6812	0.8205
Part 2	4713	0.5677	6195	0.7462	6986	0.8415	7416	0.8933	7653	0.9218
Part 3	5404	0.6509	6710	0.8082	7337	0.8838	7672	0.9241	7859	0.9466
Part 4	6020	0.7251	7118	0.8574	7590	0.9142	7859	0.9466	8017	0.9657
Part 5	6486	0.7813	7429	0.8948	7853	0.9459	8021	0.9662	8112	0.9771
Part 6	6815	0.8209	7614	0.9171	7925	0.9546	8074	0.9725	8149	0.9816
Part 7	7022	0.8458	7755	0.9341	8026	0.9668	8149	0.9816	8209	0.9888
Part 8	7282	0.8771	7857	0.9464	8079	0.9731	8171	0.9842	8216	0.9896
Part 9	7494	0.9027	7978	0.9610	8134	0.9798	8205	0.9883	8234	0.9918
Part 10	7642	0.9208	8045	0.9694	8189	0.9867	8239	0.9928	8270	0.9965

从实验结果可以看出,在排名前 10%用户这一列中,问题数目从 3368 增长到 7642,所占的比重也从 40.57%增长到 92.08%;在排名前 20%用户这一列中,问题数目从 4957 增长到 8045,所占的比重也从 59.71%增长到96.94%;排名前 30%用户这一列中,可看出问题数目从 5825 增长到8189,所占的比重也从 70.16%增长到98.67%;排名前 40%用户这一列中,可看出问题数目从6415 增长到 8239,所占的比重也从 77.27%增长到 99.28%;排名前 50%用户这一列中,可以看出问题数目从 6812 增长到8270,所占的比重也从 82.05%增长到99.65%。通过实验数据显示,在时间跨度小的问题段,所占的比重都很高,说明高重要性的用户回答的最佳答案都基本是在较短时间内完成的,体现出了高重要性用户较高的问题回答效率。

3.6　问题推送可靠性分析

为了验证重要性排序靠前的用户解决问题能力较强,本实验将对比排名靠前用户回答问题的得分情况与其他用户的平均得分情况。

从数据集所包括的 263541 条回答信息中,选取用户重要性排名前 100 名用户的 a 类用户的所有回答,这 100 个用户回答的问题有 29907 条;29907 条问题中所有回答者为 c 类用户;这 29907 条问题中,除排名前 100 的用户回答问题以外,还有另外 26652 个用户回答了这些问题。因此,将这 26652 个用户看作 b 类用户,如表 8 所列。我们可以计算得到同样回答这些问题的所有用户平均得分、排名前 100 的用户平均得分,以及其他用户的平均得分。29907 条问题的所有回答者回答问题的平均得分是 1.6759,重要性排名前 100 用户的平均得分为 2.4075,重要性排名在 100 名以后的回答者的平均得分是 1.5435。由此可见,问题推荐给重要性排名靠前的用户回答的质量更高。

表 8　回答得分与用户重要性的关系

Table 8　Relationship between answer score and importance user

	a 类用户	b 类用户	c 类用户
The average score of Answer	2.4075	1.5434	1.6759

4　在线问答平台及案例分析

为了验证用户重要性评估方法在在线问答平台中能够应用识别重要用户,我们应用该方法开发了一个在线问答平台。该平台使用 Spring Boot 开发框架,在消息流的处理上采用基于内存处理的 Redis 插件,在前端页面处理上使用 FreeMarker 搜索引擎,采用 MyBatis 实现代码中相关配置的加载,以及结构化查询语言的解析和执行,用 Solr 搜索应用服务器搜索问题等关键字,用 Mysql 来存储平台数据。通过本文提出的基于用户关系网络的用户重要性评估方法来量化用户的重要性,且利用 MapReduce 使得算法的计算效率得到大幅提升。在线问答平台的首页如图 6 所示。

图 6　在线问答平台首页

Fig. 6　Home page of Q&A platform

4.1　平台主要功能

我们开发的在线问答平台的整体架构如图 7 所示,主要功能如下。

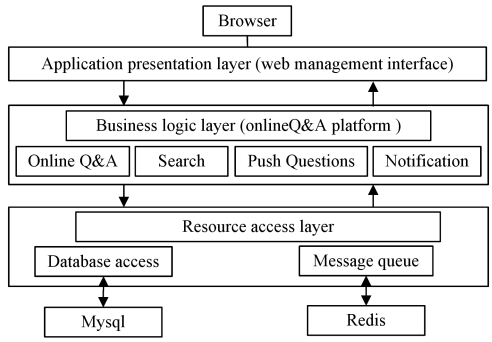

图 7　在线问答平台架构图

Fig. 7　Architecture of Q&A platform

(1)问题发布。在线问答平台首先实现用户发布问题的功能,在平台展示用户想要获知的问题,以供其他用户浏览。平台问题提出的界面如图 8 所示。

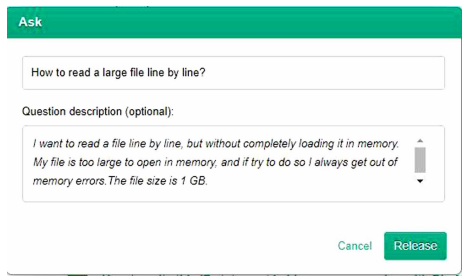

图 8　在线问答平台用户提出问题功能

Fig. 8　User submit question function of Q&A platform

(2)问题回答。用户发布问题以后,希望得到问题的答案,因此平台中的问题被其他用户浏览以后可以供其他用户或者用户自己进行回答。平台用户对问题的回答如图 9 所示。

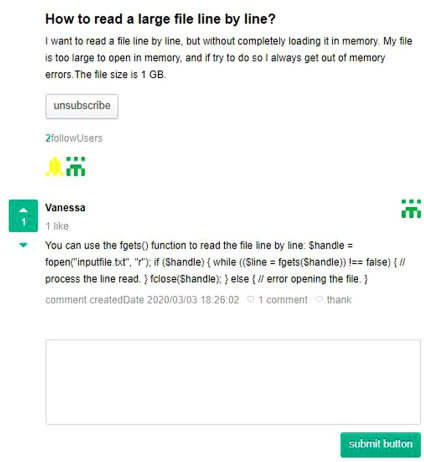

图 9　在线问答平台问题详情

Fig. 9　Question details of Q&A platform

(3)用户排名。用户通过发布和回答问题后形成复杂的用户关系网络,通过这些关系计算出用户的重要性,并对用户进行排名,为激励用户和问题推送提供依据。如图 10 所示,score 是根据用户的问答关系网络计算的用户重要性指标,由排名可知用户 Hayley 是个重要性最高的用户,其回答问题的部分信息如图 11 所示。平台对所有用户都进行了排名,这里

只仅截取了重要性排名前 20 的用户。

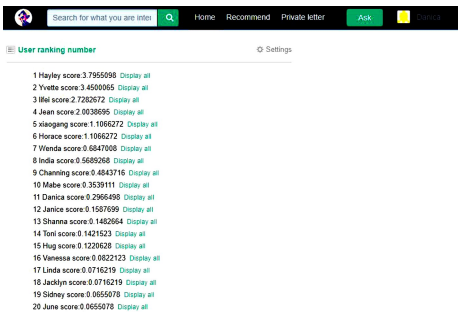

图 10 前 20 名重要用户排名

Fig. 10 Importance ranking of top 20 user s

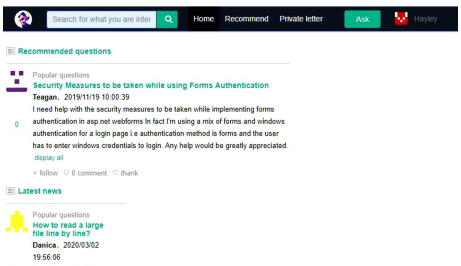

图 11 重要用户 Hayley 的信息界面

Fig. 11 Interface of High-importance users

除此之外，在线问答平台还具有根据用户重要性进行问题推送，以及对感兴趣的问题进行关注、用户间发送私信、消息通知和查看等功能。

4.2 重要用户识别验证

从图 10 中可以看到平台目前排名前 20 的用户列表，其在平台中提问及回答问题的统计数据如表 9 所列。

表 9 排名前 20 的用户数据分析

Table 9 Data analysis of top 20 users

The number of user importance	User name	The count of answers	The count of Question answered	Few statistics, relatively high ranking
1	Hayley	6	51	No
2	Yvette	7	59	No
3	lifei	0	6	Yes
4	Jean	11	31	No
5	xiaogang	1	2	Yes
6	Horace	1	2	Yes
7	Wenda	7	7	No
8	India	2	2	Yes
9	Channing	6	6	No
10	Mabe	9	7	No
11	Danica	18	5	No
12	Janice	13	3	No
13	Shanna	13	2	No
14	Toni	11	2	No
15	Hug	1	1	Yes
16	Vanessa	14	2	No
17	Linda	14	1	No
18	Jacklyn	15	1	No
19	Sidney	10	1	No
29	June	9	1	No

我们可以看出，这些用户中既包括统计数据显著的重要用户，如排名第一的 Hayley 等，也包括一些统计数据不显著的用户，如用户 lifei，其回答数为 0，问题被回答数为 6，统计指标相对于其他人要少许多，但是它的排名却比较靠前。通过查看用户 lifei 的提问回答记录可以看到，与用户 lifei 存在问题回答关系的用户是 Hayley，Yvette，Jean 等这样的重要用户，因此它的排名也相对较高。也就是说，通过我们的用户重要性评估方法，可以从在线问答平台中识别出通过统计指标不易发现的重要用户，即识别重要用户更加准确合理。

结束语 本文提出一种在线问答平台用户重要性评估方法，根据用户提问和回答行为，进行用户关系网络建模，在此基础上计算用户重要性并排序。为了验证方法的有效性，采用 Stack Overflow 数据集进行实验分析，结果表明该方法得到的用户重要性与用户在实际中的表现相符。通过分析随机删除重要性排名较高的用户后用户关系网络的破碎程度可知，重要用户在用户关系网络构成中起着重要作用；通过用户重要性排名与问题的难易程度、问题回答数、回答得分、回答的及时性等指标相对关性进行分析，证明了用户重要性排序结果的合理性；最后通过问题推送分析实验证明了对重要用户推荐问题可提高问题回答的效率。

应用用户重要性评估方法，利用 Spring Boot 框架实现了一个在线问答平台。用户可以进行提出问题、回答问题、私信、消息通知等基本在线问答平台的操作外，平台还提供了重要性排名、问题推送等功能。案例分析表明，依据用户重要性排名进行问题推荐，能够加快问题的解决速度，提升用户的体验感。

另外，本文提出的在线问答平台中的用户重要性评估方法除了可以应用到在线问答平台，还可以应用于其他系统，如社交网络中用户的重要性评估及高影响力用户的识别等。

本文还存在一些不足之处：对用户重要性计算应用的关系可能考虑得还不够全面，未考虑与用户属性有关的其他因素；在线问答平台中的问题在推送时，在低质量问题的识别上考虑的因素过少。未来工作将针对以上不足开展。

参 考 文 献

[1] YE D H, XING Z C, KAPRE N. The structure and dynamics of knowledge network in domain-specific Q&A sites: a case study of stack overflow [J]. Empirical Software Engineering, 2017, 22: 375-406.

[2] LI L, HE D Q, ZHANG C Z. Survey on Social Question and Answer [J]. Data Analysis and Knowledge Discovery, 2018, 2(7): 1-12.

[3] GUO J W, XU S L, BAO S H, et al. Tapping on the potential of Q&A community by recommending answer providers [C] // ACM Conference on Information and Knowledge Management. 2008: 921-930.

[4] ZHANG Z F, LI Q D. Studies on Community Question Answering—A Survey [J]. Computer Science, 2010, 11(11): 19-24.

[5] ZHAO Y X, PENG X X, LIU Z Y, et al. Factors that affect asker's pay intention in trilateral payment-based social Q&A platforms: From a benefit and cost perspective [J]. Journal of the Association for Information Science and Technology, 2019, 71(5): 516-528.

[6] ZHANG K H, LIANG J Y, ZHAO X W, et al. A Collaorative Filtering Recommendation Algorithm Based on Information of Community Experts [J]. Journal of Computer Research and Development, 2018, 55(5): 968-976.

(下转第 448 页)