Respond Rank: Improving Ranking of Answers in Community Question Answering

Geerthik S[1], K. Rajiv Gandhi[2], Venkatraman S[1]
[1] Computer Science, PRIST University, India
[2] Alagappa University Constituent College, India

Article Info

Article history:

Received Nov 29, 2015
Revised Jun 24, 2016
Accepted Jul 8, 2016

Keyword:

CQA
Ranking techniques
Respond Rank

ABSTRACT

Ranking is used in Community Question Answering (CQA) for positioning user answers. Different ranking techniques are used in CQA for ranking user answers. We identified three drawbacks with the existing ranking. The Quality answers written recently are not recognized properly compared to old average answers. Also the answers from the users having less number of followers are not recognized properly compared to users with more number of followers. Moreover experts and normal user likes are treated equally. We propose RespondRank for ranking the user answers. RespondRank identifies best answer better compared to existing methods.With RespondRank Quality answers from different users are recognized properly. Experiments carried out on Quora a popular CQA, shows our RespondRank shows significant improvement in ranking than the existing ranking techniques.

Corresponding Author:

Geerthik S,
Research Scholar Computer Science,
PRIST University,
India.
Email:

1. INTRODUCTION

Community question answering (CQA) is a type of information retrieval were user need from the community is given in the form of natural language question and community response is in the form of natural language answer. The great success of CQA leads to programming wizards like stackoverflow.com [1], Quora a place for knowledge sharing [2],[3], yahoo answers [4], zhihu [5] etc. Many users register with CQA and only a few are active. Analyzing a Java forum [6] with social network analysis method, the results show that only 12 % of users ask questions and answer the questions among themselves. They are only the highly active user . Also only 13 % of users only answer questions and nearly half of the users only ask questions. Quora contain nearly 100 million users were 80 % of people are only visitors who read the answer content. Only 20 % of users vote the answer and follow people. In majority of CQA's ranking is based on only a small crowd of users.

New challenges for researchers are created by CQA sites. Some of the challenges we identified in CQA systems are identification of experts, answer ranking, dealing with unformatted questions, dealing with unanswered questions. The users who are writing the answer to the questions may be expert in the subject he is answering or new to the particular topic. As a result, most of the CQA sites fail in the quality ranking of answers. So the most important challenge in CQA is discovering a quality answer from group of answers written by different users [7].

Considering the questions, they are posted by users in different domains. The quality of the question can be measured with the number of upvotes and downvotes for the question, question tag, length of the question and the number of answers received to the question [8]. Different types of questions posted in CQA are definition, opinion, procedure, reason, factoid,Why,Yes/No questions. Among these factoid and

Yes/No questions are difficult to answer because they need more facts and explanation in answering [9]. Also, the factoid, opinion and definition questions were attracting more users to write answers than other types of questions [10]. Factoid answers are answered with the help of opinions which is based on persons desire,belief or speculations [11]. Also Why question is explained by user only with the help of material,person or a purpose [12].

Considering the answers, the answer written to the question must be relevant to the question. Some quality answers have an introductory part, plot and theme of answer. The user must be patience in writing the answer and if needed the answer should be descriptive. Some answers need some proofs and also some images. So the best answer must attach the images or the URL link for the references if needed. Some question only needs a factual answer and for such type of questions the answer must be short and clear. Considering the users the best users in CQA have three properties. They must give timely response, they provide quality answers, and they maintain a good social profile [13].

This paper is organized as follows: section 2 discusses the problems with the current ranking system, section 3 describe about the related work in ranking CQA, section 4 describes about RespondRank algorithm in ranking, section 5 gives the results and discussion, the future work and main conclusions are given in the last section.

2. PROBLEMS WITH RANKING ANSWERS

CQA sites are creating archives of hundreds of questions and millions of answers every day. So identifying the best answer is a basic need in CQA sites. Different factors used by CQA sites for ranking the answers are, total upvotes to the answer, total downvotes to the answer, previous answers written by the author, whether the author is an expert in the subject or not and quality of the content, etc. With the factors mentioned above the answers with more upvotes are ranked higher and the answers with more downvotes are ranked lower.

In any CQA site people will be reading only top two or three answers and upvote any one of the answers. The answers which are not initially written and in the bottom are not read and considered by many users. One method in ranking to overcome this problem is to hide the likes for the users for all answers for some period of time. If a user answer got very good upvote compared to other user answers, then the answer is declared as the best answer and then upvotes are visible to all users . But practical implementation of this method is difficult.

A basic statistical method of ranking is, identifying the users who answer more on a particular topic and consider him as an expert and give his upvote more importance. But this is not true, in the case of online advertisement and spammers who are intended in promoting their products online.

The rest of this section gives the background analysis of followers in CQA ,relationship between the number of followers and answer view and also discuss the two major problems we identified in the CQA system.

2.1. Background analysis on follower in CQA

Followers on CQA for any user depends on the quality of his answer, blogs he maintains, whether he is a celebrity or not, topic of expertise, gender, etc. At the same time, the number of followers also depends on the total time a user is spending on CQA sites. Consider a user Alex K. Chen who asked 45,113 questions in Quora. He had more than 12,000 followers. Consider another user Marc Bodnick who asked 12,591 questions and he had more than 61,391 followers. On the other hand, 60% of users in Quora had less than 10 followers. If a new user register to CQA sites, the number of followers, he got will be very less.

2.2. Relationship between number of followers and answer view

The total number of upvotes also depends on total number of answer views. Answer view is the option by which user can read one or two lines in every answer. If he is interested in reading the entire answer, he can view the entire answer. Answer view is more for the answers which are written by users who contain many followers. Consider a user Joel, who had written an answer for a question. Consider if another user Scott who follow Joel, upvotes Joel answer. Now the Joel answer is visible to Scott followers. If the followers of Scott likes Joel answer they upvote it. If Scott is having many followers the answer views is even increased. It is clear that there is a strong relationship between the number of followers and number of answer views.

2.3. Problem 1: Problems faced by new users

In most of the CQA sites like Quora, Stackoverflow any user can contribute the answer to the question. But most of this site is very unfair to the new users and the answers provided by them. The question and answers users read in their timeline are from the people and stuff they follow [5],[13]. The ranking of answers is mostly determined by the number of followers for the user who writes the answer and not the quality of the answer. So most of the answers ranked tops are from the users who have more followers even though the answers may not be interesting sometimes. In other words, if a new user with few followers, writes a good answer, his answer is not ranked in top. The good answers are not reached to the maximum users due to this problem.

Topic categorization [14] is found in most of the CQA sites which help users mainly to reduce the time in searching answers. But if quality answers are not ranked in top users need to spend more time in reading all the answers for that particular question for getting the needed knowledge. Many users write the answer in CQA sites for some personal happiness they got from the upvotes. So many new users were not willing to write answers in the CQA sites.

2.4. Problem 2: Problem faced by Late Quality answers

To illustrate this problem, consider the Table 1, where a user asks a question in the bio-medical field.

Table 1. Week vrs Rank

Users	Week	Rank obtained
UserA1	Week 1	Rank 1
User A2	Week 1	Rank 2
-	-	-
User A50	Week 1	Rank 50
User A51	Week 2	Rank 51

Let's consider if *UserA1* answer the question in the first week itself and it is voted by many users as the best answer from a list of 50 answers. Consider if *UserA51* answers the same question after one week and the answer is better compared to *UserA1* answer. Since the answer is written after one week initially the answer is displayed in 51 position. In most of the CQA sites many users read only top few answers. After reading two or three top answers, they vote any one of the top answers listed at that time. So in most of the cases, answer written by *UserA51* is not read by many users. Even though *UserA51* answer is more quality answer than the *UserA1* answer it is not visible to all users. In our RespondRank method this problem is removed by using answer view as one of the feature in the ranking system.

By using the percentage of answer views and percentage of follower upvote in our ranking method the problems given above are greatly reduced by RespondRank algorithm.

3. RELATED WORK

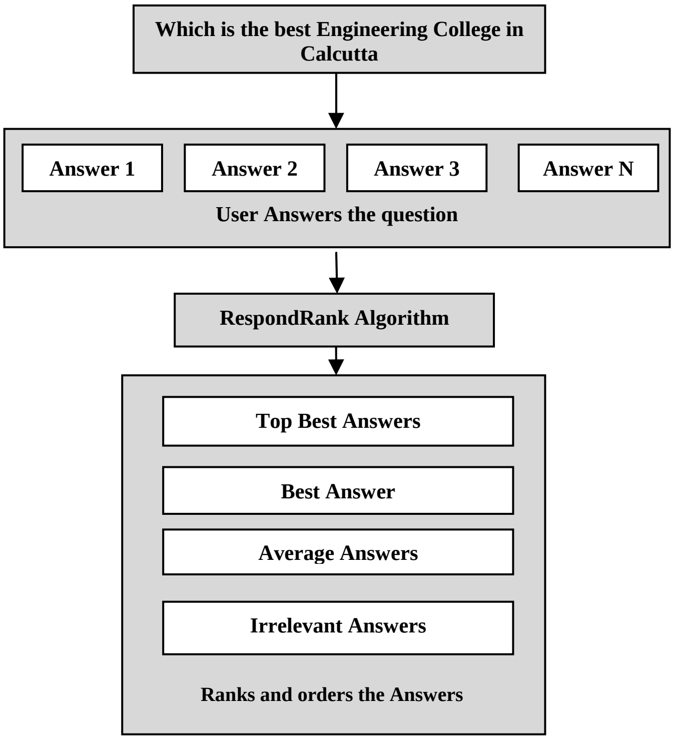
A hybrid hierarchy-of-classifiers framework for finding the quality answers in yahoo answers is proposed by [9]. Before analyzing the different answers to a given question, the question is analyzed first. The user answers are compared with the expected answers which are already stored for the different question types. The framework is compared with different questions from yahoo answers and the best answer prediction is very good. Usage of z-score measures in identifying the experts is done in [6] were if the users ask and reply equal number of times on a particular topic his z-score is 0. If the user answer more than the questions he posted for a particular topic his z-score is positive value else if the user ask more on a particular topic and answer fewer his z-score is negative. The problem with z-score algorithm is this algorithm only based on the number of replies and not based on good replies on particular topics.

An analogical reasoning approach for ranking, where ranking for the answers is based on the resemblance between new question answers and existing best similar question answers [15]. If a user writes an answer for the given question, the question and answers are compared with the existing best question answer of similar type and ranking is done. This method is good for finding the textual mismatches and answers spam.

Most of the CQA uses page rank algorithm for ranking answers. Consider three users X, Y and Z. With this page rank algorithm, if the X answer to Y and Y answers to Z then X is considered an expert among X, Y and Z. This is an efficient method for predicting the experts in CQA sites.In like manner [13] uses AA page rank algorithm with Quora data sets and calculated authority score and activity score of users in identifying potential answer supplier. The authority score increases with reply count and it decreases with question count. Activity score is calculated based on the frequency of user visiting CQA. Here the user upvote with high authority and activity score get more importance. Accordingly Zhihurank [5] identified the user authority in ranking based on link structure and topic similarity between question and his expertise field. In this method if User X with high authority and User Y with medium authority upvote Z, then upvote from user X is treated more powerful than user Y which increase authority of user Z.

4. RESPONDRANK ALGORITHM

We can apply the RespondRank algorithm in answer ranking, best answer finding, and positioning of answers. The Figure 1 gives the overall architecture of CQA sites. Here, a user posts a question and other users answer the question. For ranking different answers and positioning them in an order, we use RespondRank algorithm. The best answers are listed in the top, followed by average answers followed by irrelevant answers.


Which is the best Engineering College in Calcutta

Answer 1

Answer 2

Answer 3

Answer N

User Answers the question

RespondRank Algorithm

Top Best Answers

Best Answer

Average Answers

Irrelevant Answers

Ranks and orders the Answers

Figure 1. Architecture of ranking in CQA

To illustrate RespondRank algorithm, consider Table 2.

Table 2. Upvote Analysis

Person	James	Justine
Total followers	2000	500
Upvotes from followers	500	180
Upvotes other than followers	200	140
Answer views from normal users	250	160

Let's assume a question is posted and it is answered by two people James and Justine. Assume person James has 2000 followers and person Justine has 500 followers.

Assume the answer by James got 753 upvotes were 500 upvotes are getting from the followers, 200 upvotes are getting from normal users and three expert votes. Also, consider answer by Justine got 324 upvotes were 180 upvotes are getting from his followers and 140 votes are getting from other people and Justine got four expert up votes. With James and Justine having total up votes 753 and 324 up votes respectively, most of the CQA sites rank the answer provided by James as highest rank and the answer provided by Justine is ranked below the James.In the above Table 2, percentages of upvotes is calculated as given.

$$\text{Percentage of upvote from followers} \quad = \frac{\text{Upvotes from followers}}{\text{Total number of followers}} \times 100$$

From table 2 the percentage of follower upvotes of James and Justine are 25 percent and 36 percent respectively. It is clear that Justine got 11 percent higher than James. Considering the above scenario we understood that most of the upvotes got from James are from his followers. Also, if there are more followers of Justine than James then, surely Justine will get more upvotes than the James. Also, the number of upvotes Justine got from normal users is higher compared with James. The expert upvotes of Justine got is also higher compared with James.But Justine answer has ranked lower than James with existing ranking methods. Let us calculate percentage of answer view of James and Justine from Table 2.

$$\text{Percentage of answer view} = \frac{\text{Total Upvotes from normal users}}{\text{Total number of answer views}} \times 100$$

James had got 200 upvotes from 250 answer views and Justine had got 140 upvotes from 160 answer views. Applying above formula *percentage of answer view* of James and Justine is 80 percent and 87.5 percentage respectively. Most of the existing ranking methods only count the total upvotes which makes the ranking flawed. In our RespondRank ranking is based on *percentages of follower upvotes* and also the *percentage of answer view* which makes ranking better.

4.1. Respond Rank definition

The RespondRank is calculated as follows

$$\text{RespondRank(U)} = \left(\frac{p}{q} \times 100\right) + \left(\frac{r}{s} \times 100\right) + \gamma + \alpha - \beta - d$$

p is the total number of upvotes from the followers
q- total number of followers..
r- total number of upvotes from normal users other than followers .
s- total answer views from normal users other than followers.
α- number of expert upvotes
β- number of expert downvotes .
γ- special upvotes, that is the number of upvotes got from other users who already written answer to the same question.
d- The total number of downvotes from normal user.
Where *U* is the user who writes the answer,

The number of expert upvotes is marked as α , total number of downvotes from experts is denoted by β. Most of CQA have some potential users who are marked as experts. The experts are identified with some parameters like users who wrote the answer in same topic already and got more upvotes. Consider a scenario where six users answers to a given question in the field of computing. Among them four answers appear to be correct for any user who read the answers. The best answer among the four is understood only by an expert in computing.

In some cases the user answer a question and he got some upvotes and if he finds some other answer is more interesting than his answer or if he likes some other answer he will be able to upvote that answer. That upvotes are special upvotes and they are denoted by γ.

It is a very rare situation the answer gets the downvotes. But if the answer is totally irrelevant to the current question, some user will downvote the answer. In our method we took the value of upvote, downvote ,expert upvote and expert downvote as one. But in real-time implementation, we can increase the expert upvote and downvote value based on the quality of expert.

4.2. Computing RespondRank

Consider Table 3 where details of 2 users are given. The Respond Rank is given by

$$\text{RespondRank(U)}=\left(\frac{p}{q}\times 100\right) + \left(\frac{r}{s}\times 100\right) + \gamma + \alpha - \beta - d$$

$$\text{RespondRank(James)}=\left(\frac{500}{2000}\times 100\right) + \left(\frac{200}{250}\times 100\right) + 3$$

RespondRank(James)=108

Table 3. RespondRank Calculation

Person	James	Justine
Total followers	2000	500
Upvotes from followers	500	180
Upvotes from normal users	200	140
Answer views from normal users	250	160
Expert upvotes	3	4
Expert downvotes	0	0
Special upvotes	0	0
RespondRank	108	127.5

Also the RespondRank of Justine is calculated similarly

RespondRank(Justine)=$\left(\frac{180}{500} \times 100\right) + \left(\frac{140}{160} \times 100\right) + +$

RespondRank(Justine)=127.5

Here Justine got higher rank than the James and he is ranked top. With this ranking method we can calculate the rank of any number of user answer participating in the discussion and position the answers. The RespondRank is any positive number, if the answer contains more dislikes or down votes then, rank is negative. The advantage of this method over existing approaches is this ranking method treats all the users under the same criteria, only the quality answer will be in the top positions.

In RespondRank method an answer is only marked as top answer if it contains high *percentage of follower upvotes* and also it contains a high *percentage of answer view*. These are two features we used in finding RespondRank and some other features like the size of the answers is not taken into consideration because some answer with a short description will be able to satisfy the user who asked the question. Most of the CQA ranking is also based on the size of the answer. But some user writes smaller quality answer that is not ranked top due to this problem. If the rank of two users is the same then, we consider the size of the answer also for determining the top rank. Some other facts considered if the answering of two users is the same are the posting date of the answer, effective use of multimedia in answer etc.

5. RESULTS AND DISCUSSION

We collected data-sets from Quora, where the average turnaround time of the user to get answers to his question is four days. We computed RespondRank for 600 questions and around 9000 answers. The general statistics of the data-set are given in Table 4.

Table 4. The general statistics of the dataset

Topics	Number of questions	Number of answers	Number of top answer differ from respondrank
#Ethnic and cultural differences	50	456	6
#Philosophy of everyday life	87	1024	9
# Minimalist lifestyle	12	86	3
# Writers and authors	53	436	7
#Starbucks	46	502	5
#Snacks	23	412	2
#Health	86	1350	8
#Higher education	68	956	8
#Life	79	1635	7
#Human behavior	32	632	3
#Working out	64	1236	7
Total	600	8725	65

Most of past research in Quora is for expert identification and comparision with other CQA,no ranking is proposed for Quora.So we compared our RespondRank with existing Quora ranking and the changes we identified is given in table 5.Among these 8725 answers we evaluated with RespondRank , 65 top answers differ from existing top answers and more than 3500 answers position is changed.

Table 5. Comparision of respondrank with quora rank

Total answers evaluated	8725
Number of changes in position with existing answers	3500
Number of changes in top position of answers	65

We also observed among these 65 top answers written users, more than 50 users had less than 100 followers. We can see that existing ranking is in favour of users having greater number of followers. Also more than 40 percent of answer positions are changed that in favour of users having greater number of followers.This results shows that our RespondRank algorithm identifies most comprehensive, trustworthy answers compared to the existing ranking methods.

The next stage of our research is based on the data-set of users having more than 2000 followers. We selected 25 users in Quora who had more than 2000 followers. The analysis is done on the past answers written by them. We selected 100 different answers written by these 25 users and we found that 68 answers among 100 answers are selected as top answers. From this it is clear that if the user have a large number of followers then, probability of getting his answers as top answer is very high. From above two results it is clear that our RespondRank method is helpful for finding quality answers written by any number of users.

6. CONCLUSION AND FUTURE WORK

This paper lists the problem faced by users in CQA sites and uses a RespondRank method for ranking answers. The main problem we identified in CQA sites is quality answers written by new users are not properly recognized. In our Respond Rank method above problem are solved by introducing percentage of follower up votes, percentage of answer view as new ranking parameters. The expert up votes and downvotes are also taken into consideration in our ranking method which increases the reliability of our ranking method. The RespondRank algorithm is evaluated with Quora data sets and it identifies the best answer better than existing methods. In our experiments we applied RespondRank algorithm for finding the best answer and positioning user answers. We can modify RespondRank algorithm for ranking the questions in CQA also. There is a need to identify the best question because some experts only answer fewer questions due to their time availability and if the CQA provides facility to identify best questions, they can answer that specific question.

REFERENCES

[1] A. Anderson, *et al.*, "Discovering value from community activity on focused CQA sites: a case study of stack overflow," *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM*, 2012.

[2] C. Rughinis, *et al.*, "Computer-supported collaborative accounts of major depression: Digital rhetoric on Quora and Wikipedia," *2014 9th Iberian Conference on Information Systems and Technologies (CISTI)*, 2014. doi:10.1109/cisti.2014.6876968.

[3] R. Rughinis, *et al.*, "Computer-supported collaborative questioning. Regimes of online sociality on Quora," *2014 9th Iberian Conference on Information Systems and Technologies (CISTI)*, 2014. doi:10.1109/cisti.2014.6876946.

[4] L. Bowler, *et al.*, "I know what you are going through": Answers to informational questions about eating disorders in Yahoo! answers: A qualitative study," *Proceedings of the American Society for Information Science and Technology*, vol/issue: 50(1), pp. 1–9, 2013. doi:10.1002/meet.14505001057.

[5] X. Liu, *et al.*, "ZhihuRank: A Topic-Sensitive Expert Finding Algorithm in Community Question Answering Websites," *Lecture Notes in Computer Science*, pp. 165–173, 2015. doi:10.1007/978-3-319-25515-6_15.

[6] J. Zhang, *et al.*, "Expertise networks in online communities: structure and algorithms," *Proceedings of the 16th international conference on World Wide Web. ACM*, 2007. http://dx.doi.org/10.1145/1242572.1242603.

[7] Z. M. Zhou, *et al.*, "Exploiting user profile information for answer ranking in cQA," *Proceedings of the 21st International Conference Companion on World Wide Web - WWW '12 Companion*, 2012. doi:10.1145/2187980.2188199.

[8] A. Baltadzhieva and G. Chrupała, "Question Quality in Community Question Answering Forums," *ACM SIGKDD Explorations Newsletter*, vol/issue: 17(1), pp. 8–13, 2015. doi:10.1145/2830544.2830547.

[9] H. Toba, *et al.*, "Discovering high quality answers in community question answering archives using a hierarchy of classifiers," *Information Sciences*, vol. 261, pp. 101-115, 2014. http://dx.doi.org/10.1016/j.ins.2013.10.030.

[10] Chua, *et al.*, "Measuring the effectiveness of answers in Yahoo! Answers," *Online Information Review*, vol/issue: 39(1), pp. 104-118, 2015. http://dx.doi.org/10.1108/oir-10-2014-0232.

[11] H. Akkineni, *et al.*, "Online Crowds Opinion-Mining it to Analyze Current Trend: A Review," *International Journal of Electrical and Computer Engineering*, vol/issue: 5(5), 2015.

[12] A. E. Karyawati, *et al.*, "Ontology-based Why-Question Analysis Using Lexico-Syntactic Patterns," *International Journal of Electrical and Computer Engineering*, vol/issue: 5(2), pp. 318, 2015.

[13] H. Wenwen, *et al.*, “Ranking potential reply-providers in community CQA system,” *Communications, China*, vol/issue: 10(10), pp. 125-136, 2013. http://dx.doi.org/10.1109/cc.2013.6650325.

[14] W. Wang, *et al.*, “Improving question retrieval in community CQA with label ranking,” *Neural Networks (IJCNN), The 2011 International Joint Conference on. IEEE*, 2011. http://dx.doi.org/10.1109/ijcnn.2011.6033242.

[15] X. Tu, *et al.*, “Analogical reasoning for answer ranking in social CQA,” *IEEE Intelligent Systems*, vol/issue: 27(5), pp. 28-35, 2012. http://dx.doi.org/10.1109/mis.2010.130.