VOL. 12, NO. 1, JANUARY 2017 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

CATEGORY BASED EXPERT RANKING: A NOVEL APPROACH FOR EXPERT IDENTIFICATION IN COMMUNITY QUESTION ANSWERING

Geerthik S.[1], Rajiv Gandhi K.[2] and Venkatraman S.[1]
[1]Department of Computer Science, PRIST University, India
[2]Alagappa University Constituent College, India
E-Mail:geerthikcs@gmail.com

ABSTRACT

The quality of the Community Question Answering (CQA) depends on the way the experts are identified from the users in CQA sites. The traditional expert identification methods struggle to identify the quality experts. Also, most of the questions routed by CQA to the experts for answering are not able to answer by them. Also, onlyvery few experts are identified by traditional CQA, the experts are not identified in many categories. This paper proposes Category Based Expert Ranking (CBER) for identifying experts in the CQA sites. We done experiments with the CQA Quora and our approach got considerable improvement over expert identification compared with the traditional methods on the multiple metrics.

Keywords: CQA, CBER, expert identification, question routing.

INTRODUCTION

Community Question Answering (CQA) gives an open social forum for users to ask questions and get a variety of answers from different set of users. The Quora, Yahoo Answers, Zhihu [4] ,Stack overflow [1] are the most popular question answering. The common challenges these CQA facing are expert identification [11], ranking the user answers [9], retrieving the quality questions [12], dealing with unanswered questions, dealing with unformatted questions [2], the low participation rate for answer by the users [14] etc.

Considering the challenges, the most important challenge is ranking of the user answers and identification of the experts. The quality of CQA gets increased if the experts are identified better. In this paper a better expert identification technique is proposed for CQA sites.

The user in CQA sites can post questions, answer a question, like an answer or route any question to other users. But compared with these activities most of the users in CQA are only interested in reading others answers. Very few users in CQA post questions and very few users in CQA are writing the answers. Also by comparing the number of questions posted and number of questions answered only a few questions are answered [6]. The different types of questions posted in CQA are *procedure, factoid, opinion, yes/no, Definition* and *reason* questions [9]. Among these questions most users are interested to answer *opinion and factoid* questions.

Since the questions, they have been posted in quite a lot of categories. Some categories like health, meditation each user had something to write answer because they had some knowledge in the basic categories. Other categories like economics, schooling the same users aren't in a position to reply. Only the specialist, experts in distinctive categories can able to put in writing satisfactory answers in such categories. Considering the CQA Quora lots of the questions posted general subject matters are already answered. However, the majority of questions in some specific categories remains unanswered. Also, in most of the CQA most effective 20 percent of users give 50 percent of excellent answers [14]. Such users have got to be recognized as the expert users. Considering that the posted answers many of the quality answers in CQA had an introductory phase, theme of answer, references, URL links and attachment of multimedia contents.

If the expert users are not identified properly, then the garbage of low quality content will probably be accumulated and it leads to failure of the particular CQA [8]. Currently the experts in CQA are identified in one of the most following approaches. In the first approach new question is routed to the user based on the previous answer he has written. In the second approach the experts are recognized based on the profile he is keeping. It involves the previous answers, past questions, previous likes for his reply, his visiting time to the CQA website, and so on.

Existing research in CQAhas achieved a promising performance in expert identification. However, we identified some drawbacks with the current expert identification methods: Most of the users writing an answer in CQA will be an expert in a minimum any one of the categories. But currently only a few users are identified as experts in each category by current expert identification algorithms. It leads to most of the questions in CQA are unanswered. Current expert identification methods use question routing techniques [3] where question is routed to the expert users for answering. Here only a few routed questions are able to answer by the identified experts. Most of the questions routed to experts for answering are unable to answer by them. Also, another drawback we identified is currently experts are identified in only a few categories. The experts are not identified in most of the technical categories. Also the expert identification fails to identify the experts in the topics which are trending. The unanswered question archive is growing day by day in CQA due to above drawbacks.

In this paper, we propose Category Based Expert Ranking (CBER) algorithm for identification of experts from all registered users in CQA in all the categories. The main objective of this paper is to identify more number of quality experts in CQA. The next objective of this research

is to identify every user who writes answersin CQA as an expert in any one category. When these two primary objectives are satisfied, the number of unanswered questions in CQA will be reduced and also the askers will get quality answers.

With CBER when the experts are identified the expert ranking isdone with each category. With CBER the knowledge level of experts in each category is identified. The rest of the paper is organized as follows: section 2 discusses the way the experts are identified in CQA and also a problem with question routing in CQA, section 3 describes CBER algorithm in ranking the experts, section 4 gives the results and discussion and the main conclusions are given in the last section.

Current expert identification in CQA

Existing expert finding existing methods is based on building an expert identification model based on past question and answer activity of users in CQA. The model is constructed based on authoritative information about users and the document relevant to the question asked. With the constructed model the experts are identified.

Many expert identification and the future expert prediction algorithmuse the votes from other users for a particular user and his past answers as the main criteria in identifying him as an expert.In addition to the past answers, the total number of answers, answer status, author reputation and content quality are some other parameters used in identifying the experts [5].

Many CQA sitesuse the page rank algorithm [7] for the ranking experts. 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 finding the experts in CQA sites.

In like manner the Zhihu rank [4] use the link structure based on page rank algorithm and another feature topic similarity for identification of the experts. The Zhihu rank expert identification calculates authority score based on number of likes user is getting and activity score based on user active participation in CQA.

Future experts are also predicted bya semi-supervised approach with the factors textual, behavior and time [10]. If a user had more than ten best answers in expertise area and he is consistent in visiting the CQA he is predicted as expert by this approach. The experiments are done with stack overflow and the experts are predicted with high accuracy.

The combination of both authoritative information based on the link structure of users and document relevance to query based on user answer details are used for expert identification by the expert rank [14].

In some cases, the expert identification model fails with some users due to the low information quality of users registered with the particular CQA. The combination of CQA sites with social networks [15] gives more information about the users and expert identification is done better.

Previous research also includes a probabilistic model in expert identification for identifying current experts and also the users who are capable of becoming the future experts [5]. Future experts are predicted based on the fact, how the user selects the question for answering.

When the Experts are identified the unanswered questions are routed to the experts based on the profile experts are maintaining [6]. Routing of questions is done by comparing the unanswered question with the profile user is maintaining. The question routing problem is Considered as classification problem and routing is done by [13] where both question parameters like title, length is matched with particular expert details like number of answers, the number of best answers etc.

Question routing in Quora

Question routing [4] is used in most of CQA for getting answers to posted questions in a short time. The questions are routed to the users who are identified as experts. As most of the users who are identified by the CQA are not real experts, such user is not able to answer the routed question [5].

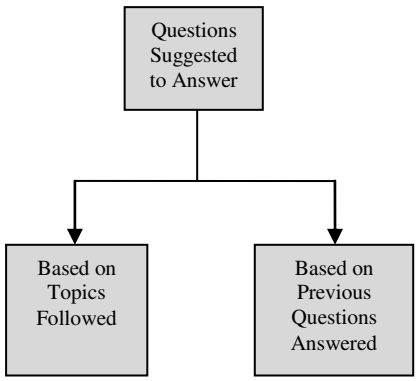

Figure-1. Question suggestion in Quora.

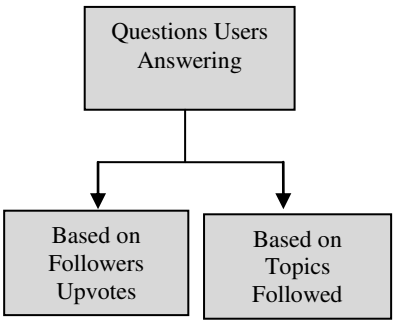

Figure-2. User answering in Quora.

Question selection by users

It is very difficult to predict the questions a particular user choose to answer it.The Figure-1tells the way Quora is suggesting questions to the users. The Question suggestion depends on the topics user is following and also based on the answers previously written by the user. But the majority of users in CQA are not interested to answer suggested questions because they are not experts in category of suggested questions.

The Figure-2 shows the way the user chooses question to answer in Quora. The questions users choose to answer are chosen from their timeline which depends on topics he followed and also based on the persons he is following. From Figure-1 and Figure-2 we infer that many users in CQA choose timeline question to answer than the suggested questions.

Relationship between question routing and expert identification

If quality experts are identified in each category and questions are routed to them properly, then most of the unanswered questions in CQA get answered in a short time. Also other problems like ranking the answers which depend on expert identification also get improved. But due to the current poor expert identification most of routed questions in CQA are unanswered.

CBER algorithm

With the CBER expertise level of all users in different category is calculated and the experts are ordered based on the expertise level. When the experts in different categories are identified well the question suggestion to the users and question routing to the users is also getting improved much better.

CBER definition

The CBER algorithm is given

Algorithm: Category Based Expertise Rank
Input: AVR, CCR, followed category, previous answers, week visit
Output: CBER for particular category

For1=1 to number of categories
$\text{CBER(category)} = AVR + CCR + Week\ Visit + \alpha \times FollowedCategory + \beta \times PreviousAnswers$
End
ReturnCBER(user)

Where
CBER-Category Based Expertise Rank,
AVR-Answer View Rank,
CCR-Common Category Relationship.
α-Followed Category Constant
β-Upvote Constant

We take the value of α and β as constant value10 in our experiments which vary based on the CQA.

After calculating the CBER for various categories for a user the top three categories with high values for the particular user are identified as the user topic of expertise. When the category rank of a user with top three categories is calculated the category rank in a particular category is compared with the different category cutoff. Based on the comparison of the category rank with category cutoff the level of expertise is determined.

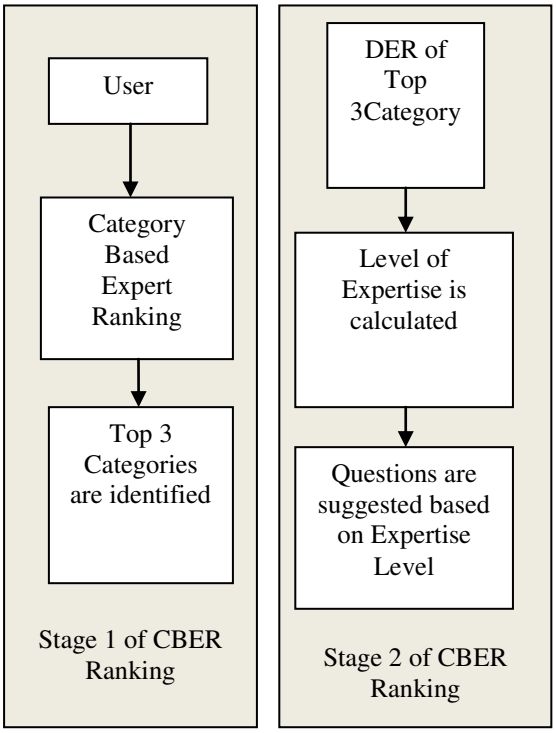

Figure-3. CBER Architectural representation.

The above Figure-3 gives the overall process in CBER. The process has two stages. In the first stage of the CBER calculation the CBER algorithm is applied to a particular user and the top three categories of his expertise is identified.

In the second stage of CBER process the level of expertise is calculated as below:

$$\text{level of expertise(category)} = \frac{\text{CBER Value of category}}{\text{Category cutoff}} \times 100$$

The percentage of the expertise of a user with a particular category is calculated with the above process.The parameters used in the calculation of CBER are given below.

Category cutoff

The category cutoff value of every category depends on the particular category. If the number of experts and the number of answers in the particular category is very high, then the category cutoff value is also high.If only few experts are available on the particular category and if the number of answers written in the particular category is also very less it makes the category cutoff value also as a smaller value. With the help of category cutoff value the level of expertise of a user in a particular category is calculated by CBER

Answer view rank (AVR)

Answer view is found in most of CQA for reading answers. The users in every field will be reading topics in the field; they are willing to become the experts. We Consider Answer View as a most important factor

inidentifying the expert because more a person is reading on a topic more he can able to answer in the particular topic. Consider a user reading following questions and answers.

Which movie or book can change your life forever?
What tiny daily habit could be life changing?
Which are the movies that inspire one to study?

After reading the answers of above questions the AVR in categories of Books, Movie, Life, Habit, Inspiration, and Study is increased by one for the particular user. Consider the same user is reading following question and answer.

Which movie can change your life?

Now the AVR of category Movie and Life is 2.By reading each question, his expertise level in particular category gets increased. Assume the AVR of user in category Movie is 6000.It means that he already read 6000 answers in the category Movie.So he can even able to answer the question in the category Movie. We take AVR valuein Consideration for identifying the experts.

Common category relationship

People in CQA get inspired by some other people in CQA and follow them. They feel that the person they are going to follow is expert in some field, they are trying to become experts. There is a strong relationship between the expertise level of a user with the people they follow. If a person follows more number of experts in a category, his expertise level is also increased in that category.

Table-1. Users and categories.

User	Categories
USER 1	Philosophy, Health, Relationship, Habit
USER 2	Health, Politics,Philosophy
USER 3	Philosophy, Health, Education, Engineering

Consider in Table-1 User 1 is following User 2 and User 3,User1 has categories Philosophy and Health in Common with User2 and User3. So we can say that User1 is interested in knowing more about Health and Philosophy and his expertise level also based on the categories health and philosophy. CCR of the User 1 depends upon the number of users the User1 is following with the common category.

For example, CCR in a category*Movie* for a user is high if a particular user follows a large number of users with expertise in category*Movie.*

CCR value gets increased with the number of users with a common category particular user already followed.

For example, CCR value is 23 for category *Movie* if particular user is following 23 users who already follow category *Movie.*

Followed category

The users in CQA follow the topics they are interested. The common topics user follows includes Health, Education etc. The experts in a particular category will be following his category of expertise for more updates. So in CBER we use *Followed Category*as a parameter in expert identification.
Consider an example

Which is best Engineering College in Calcutta?

With CBER the question will be suggested to the users to answer if they follow topics like *Engineering College* and *Calcutta.*
While calculating CBER

$$\textbf{Followed category} = \begin{cases} 1, & if\ user\ follows\ particular\ domain \\ 0, & if\ user\ not\ follows\ particular\ domain \end{cases}$$

Previous answers

Expert Identification also depends on the previous answers written by the users.

Most of the experts are identified based on the previous answers written by them. Consider the below questions

What is the use of meditation?
Do people get better health with meditation?
How to do Kriya yoga?

If a user writes answer above questions and get good likes, then he is treated as expert in meditation.

The CBER value of a category gets increased with the previous answers in the same category. If the user writes a number of answers in a particular category, his CBER value will be very high.

Week visit

Week visit is a number which depends on the number of times a user login to CQA for past one monthand spend in it.Only the past one month is taken into Consideration for calculating the Week Visit.The Week visit value is 252 if particular user spends 252 minutes in CQA for past one month.

Computing the CBER

CBER is calculated as given below:

$$\text{CBER} = AVR + CCR + Week\ Visit + \alpha \times FollowedDomain + \beta \times PreviousAnswers$$

Where
AVR-Answer View Rank
CCR-Common Category Relationship
α-Followed Category Constant
β-Upvote Constant

While calculating CBER the values of AVR and CCR values can be directly got from the CQA

Consider the Table-2 were several categories, followed by User A is given. Assume that user A is following 2000 users and his average week visit is 100 minutes.

Table-2. User adetails.

Category answers viewed	AVR	CCR	Followed s categories	Previous answers	CBER
#Ethnic and cultural differences	50	20	NO	1	180
#Philosophy of everyday life	870	560	YES	22	1760
# Minimalist lifestyle	12	45	YES	1	177
# writers and authors	123	455	NO	12	798
#Starbucks	46	22	NO	2	188
#Snacks	23	32	NO	1	165
#Health	860	353	YES	22	1543
#Higher education	68	34	NO	2	222
#Life	790	324	YES	23	1454
#Human behaviour	320	344	YES	32	1094
#Working out	64	233	YES	13	537

The expertise of a user in category*Ethnic and cultural differences* is computed with CBER as given

$$\text{CBER}(Ethnic\ and\ cultural\ differences) = AVR + CDR + Week\ Visit + \alpha \times FollowedDomain + \beta \times PreviousAnswers$$

Here
AVR=50, CCR=20, Week Visit=100, α, β=10, followed category =0, Previous answer =1

$$DER(Ethnic\ and\ cultural\ differences) = 50 + 20 + 100 + 0 + 10 \times 1$$
$$DER(Ethnic\ and\ cultural\ differences) = 180$$

In Like manner CBER of all categories are calculated.

Analysing the CBER of different categories in Table-2 the CBER values of Philosophy of Everyday Life, health, human behaviour is identified as top three categories. So we can say that the user A is expert in above categories. Even though UserAfollowed topics snacks his expertise level in that category is very low. Also the expertise level of user A in category s like working out, writing is intermediate.

Levels of experts with CBER

The below table gives the category cutoff of different categories

Table-3. Category cutoff values.

Categories	Category cutoff
Health	7897
Philosophy of Everyday Life	3567
Human behavior	4678

The CBER values of above categories from Table-2 is given
Health-1543
Philosophy of Everyday Life-1760
Human behavior-1094
The Level of expertise is given by

$$\text{level of expertise(category)} = \frac{\text{CBER Value of category}}{\text{Category cutoff}} \times 100$$

For User A

$$level\ of\ expertise(Health) = \frac{1543}{7897} \times 100 = 19.53\ \%$$

$$\text{level of expertise(Philosophy of Everyday Life)} = \frac{1760}{3567} \times 100 = 49.34\%$$

$$\text{level of expertise(Human Behaviour)} = \frac{1094}{4678} \times 10 = 23.38\%$$

From the Level of Expertise values we can easily judge user A and his top expertise categories. The Questions are routed to user A based on his Level of Expertise.

RESULTS AND DISCUSSIONS

Data sets

We collected datasets from the Quora, where the average turnaround time of the user to get answers to his question is four days. The Table-4 shows the statistics of first data set were 12000 questions are routed to 100 different users for answering and only 382 questions get answered. Only 3.18 percent of routed questions got answered. This result shows that the current expert identification method is very poor. Most of the questions routed for answering are not able to answer by the users.

Table-4. Statistics of dataset.

Total users	Total questions routed	Total questions answered	Percentage
100	12000	382	3.18

Evaluation metrics

Some standard evaluation metrics in information retrieval are taken into Consideration for checking quality category experts

Mean reciprocal rank (MRR)

The reciprocal rank of a query response is the multiplicative inverse of the rank of the first correct answer.

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i}$$

Where $|Q|$the total number of queries is, $rank_i$ is the position of the correct answer.

Precision at K (P@K)

Here K is the position of the correct answer. The precision at K gives quality answer proportion in the position K.

Table-5. Statistics of second dataset.

Total users	Expert in one category	Expert in two category	Not experts
14	5	4	5

The above table shows the statistics of the second data set. We apply CBER with 14 different users and we found that four users are expert in two categories, five users are expert in one category and five users are not identified as an expert in any one category. Totally nine experts are identified from the 14 users. With the identified experts, most of them are Level 1 and Level 2 experts. Five users are not identified as expert in any one category because they had written no answers in the past. Also with the CBER 65 percent of users is identified as experts in minimum any one of the categories, 28 percent of users are identified as experts in more than one category.

For checking the reliability of CBER the past answers written by these 14 users is taken into consideration. We got total 145 answers written by these 14 users. Also 1214 questions are suggested to these users for answering. With the questions suggested only 21 questions are answered by these 14 users. From the 21 questions answered more than 15 questions answer category matched with the identified expert category. This result shows that most of the answers the user write answers to the question category were they are experts and also the identification of the experts by CBER is really good.

For checking the quality of answer we apply MRR and precision to these 15 questions answered and we got high MRR value and low precision value. High MRR value and low precision value assures the quality of answers written by the identified experts. Also, if we extend this approach to more users we get more experts in each category. The number of experts identified by the CBER is higher than existing approach and it is needed for today CQA system. Also users are eager to answer suggested questions because experts in CBER are identified by taking answer view rank as one parameter. With CBER the experts in category can be predicted even the user not started writing his first answer.

CONCLUSIONS

Identification of experts is the main problem we identified in Community Question Answering systems. If the experts are identified properly, it leads to success of question routing also. Also, only limited experts are predicted by previous expert identification methods in some categories.

We focused on predicting category experts who can able to answer routed questions to them. We introduce CBER algorithm for identifying category experts. The experts are identified in CBER based on five different parameters. Here every user is checked for his expertise category based on the CBER parameters and most of the users we identified as experts in one category. Some users are identified as experts in more than one category. Our experimental results indicate high feasibility in expert identification. Our works indicate a promising direction towards expert identification and question routing, which play a major role in improving the quality of CQA services.

REFERENCES

[1] Allamanis M.and Sutton C. 2013. Why, when, and what: Analyzing Stack Overflow questions by topic, type, and code. 2013 10th Working Conference on Mining Software Repositories (MSR).

[2] Dror G., Maarek Y.and Szpektor I. 2013. Will My Question Be Answered? Predicting "Question Answerability" in Community Question-Answering

Sites. Lecture Notes in Computer Science. 499-514. doi:10.1007/978-3-642-40994-3_32.

[3] Li B. and King I. 2010. Routing questions to appropriate answerers in community question answering services. Proceedings of the 19th ACM International Conference on Information and Knowledge Management - CIKM '10. doi:10.1145/1871437.1871678.

[4] Liu X., Ye S., Li X., Luo Y. and Rao Y. 2015. ZhihuRank: A Topic-Sensitive Expert Finding Algorithm in Community Question Answering Websites. Lecture Notes in Computer Science. 165-173. doi:10.1007/978-3-319-25515-6_15.

[5] Pal A., Harper F. M. and Konstan J. A. 2012. Exploring Question Selection Bias to Identify Experts and Potential Experts in Community Question Answering. ACM Trans. Inf. Syst. 30(2): 1-28. doi:10.1145/2180868.2180872.

[6] Riahi F., Zolaktaf Z., Shafiei M. and Milios E. 2012. Finding expert users in community question answering. Proceedings of the 21st International Conference Companion on World Wide Web - WWW '12 Companion. doi:10.1145/2187980.2188202.

[7] Shahriari M., Parekodi S. and Klamma R. 2015. Community-aware ranking algorithms for expert identification in question-answer forums. Proceedings of the 15th International Conference on Knowledge Technologies and Data-Driven Business-i-KNOW'15.

[8] Srba I. and Bielikova M. 2016. Why is Stack Overflow Failing? Preserving Sustainability in Community Question Answering. IEEE Software. 33(4).

[9] Toba H., Ming Z.-Y., Adriani M.and Chua T.-S. 2014. Discovering high quality answers in community question answering archives using a hierarchy of classifiers. Information Sciences. 261: 101-115.

[10] Van Dijk D., Tsagkias M.and de Rijke M. 2015. Early Detection of Topical Expertise in Community Question Answering. Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR'15.

[11] Wang G. A., Jiao J., Abrahams A. S., Fan W. and Zhang Z. 2013. ExpertRank: A topic-aware expert finding algorithm for online knowledge communities. Decision Support Systems. 54(3).

[12] Zhou G., Xie Z., He T., Zhao J. and Hu X. T. 2016. Learning the Multilingual Translation Representations for Question Retrieval in Community Question Answering via Non-Negative Matrix Factorization. IEEE/ACM Trans. Audio Speech Lang. Process. 24(7): 1305-1314. doi:10.1109/taslp.2016.2544661.

[13] Zhou T. C., Lyu M. R. and King I. 2012. A classification-based approach to question routing in community question answering. Proceedings of the 21st International Conference Companion on World Wide Web - WWW '12 Companion.

[14] Zhou Z.-M., Lan M., Niu Z.-Y.and Lu Y. 2012. Exploiting user profile information for answer ranking in cQA. Proceedings of the 21st International Conference Companion on World Wide Web - WWW '12 Companion.

[15] Zhao Z., Zhang L., He X.and Ng W. 2015. Expert Finding for Question Answering via Graph Regularized Matrix Completion. IEEE Transactions on Knowledge and Data Engineering. 27(4).