Stack Overflow Technical Response Analysis for Contemporary Communities on Internet

Md Naiyer Hoda[1], Irfanul Haque[2]

[1] P.G Scholar, Department of Computer Science & Engineering, RKDF College of Engineering, Bhopal, India

[2] Ex. Asst. Professor, MIT Purnea, Bihar

Abstract - **The popularity of community question answer (CQA) forums like StackOverflow, Yahoo Answers and Quora is increasing tremendously with thousands of questions being posted each day and about thrice the number of responses being provided. With such query explosion, users participating in these forums receive a huge number of postings that adversely affects their responsiveness and also the quality of the responses. Hence, identifying topical experts is necessary to improve the efficacy of these systems in terms of both response time and quality. Although expert detection in CQA forums has traditionally been a topic of wide interest, however, many of the proposed techniques use features set that reflect the popularity of the responses of the responder rather than the difficulty level of the questions being responded. In this paper we provide measures of labeling difficult questions and use the number of difficult questions responded by a user combined with other user interaction parameters in identifying potential topical experts. Using a random forest classifier with the proposed feature set on StackOverflow data, we obtain an improvement in accuracy of 5 - 16% over existing techniques, in detecting topical experts.**

I.INTRODUCTION

Community question answer (CQA) forums like Stack-Overflow provide a platform where users can post questions on specific topics that are responded by relevant users who are possibly aware of the topic or about the specific query in the post. Although the topical community sizes in these forums are typically very large and the number of posts ranges to millions, however empirical studies indicate that a huge majority of the questions posted receive only one or two answers and many of them remain unanswered[1]. Due to the huge number of posts, the users in the community face a problem of query explosion that affects their responsiveness to the queries. Hence efforts have been made to incentivize suitable responses to the queries by providing best answer marking or badges. However, another popular line of thought to improve these systems is to identify topical experts and categorize questions based on the expertise required so that a limited set of questions are pushed to the users based on their expertise [2]. Most of these expert detection techniques use different set of features like reputation score (badges), the user Z-score [3] as well as the difference between up and down votes for all previously submitted answers [4] , while Zhou used topical similarity between questioners and answerers [5]. However, the problem still remains open and there still remains a huge scope of improving the accuracy in detecting topical experts in these forums.

In this paper, we propose identifying topical experts in a StackOverflow community by considering the difficulty level of the questions that were answered by the responders previously combined with other interaction based features like the typical score and frequency of the questions and responses posted along with the responder Z-score. We label questions as difficult based on two different aspects:

1) Questions whose solutions are not easily found, i.e. the response time is high and the number of responses are low

2) Questions that are raised by a large number of people. We consider several topical features that capture these aspects to label the difficult questions. We use a six year StackOverflow data for our analysis. The data set contains the questions, answers (posts), tags (topic), scores of posts, best answers, view count of questions, favorite count, post type id as well as timestamp of posts along with certain other parameters. We apply a random forest classification technique on the proposed feature sets and compare

the accuracy of the proposed approach for four expert labeling mechanisms, a) when the number of best answers of a responder is above a threshold, b) when the number of difficult questions answered is above a threshold, c) when the best answers count as well as the answers in low response questions is above a threshold and d) when the best answers and the number of difficult questions answered is above a threshold.

We empirically show that although the difficulty level of the questions that are responded by the users may not be the only single best feature for identifying experts, however by combining this feature with other relevant features based on the user interaction pattern, the accuracy of the expert detection technique improves drastically. We have validated our approach on three different programming language based topics and the improvement in accuracy varies from 5 - 16%.

The organization of the paper is as follows. Section II provides a review of the prior work on the topics concerned. In section III, we describe in details our proposed approach of topic detection and the feature sets used. Section IV provides an overview of the experimental procedures and analysis of the results obtained. Section V highlights the conclusion of our work.

II. LITERATURE REVIEW

In this section we provide a brief review of the existing works in the relevant area. Certain works exist in the area of identifying difficult questions. Lin [6] uses 'knowledge gap based KGD-rank algorithm' approach to identify the difficult questions. Unlike this probability based approach that requires extensive analysis of the user-user network, we follow a simple semi supervised machine learning approach based on the question and response features that are available directly in the data. Moreover, our basis of categorizing depends on the generalized features of difficult questions. Liu proposed an approach to include question quality in determining the answer quality in a CQA services [7]; however, apart from the question quality, we also consider additional features like number of comments, answerers Z-score etc., in labeling the difficult questions.

There have been several works on expert identification. In [8], the authors presented the work of two statistical topic models to find relevant experts for the recent posted questions and provided a better approach than LDA. Cong et al. presented the similarity in question and answer pairs using graph based approach. The work helped in establishing transitive association of the questioners and the answerers [9]. This is more likely to find an expert for the question type. Similarly, another work of question and answer pairs similarity is given in [10]. Wang et al. used textual features and the user's authority in community by a modified page rank technique for identifying experts [11]. We have used a similar document based method as used in this paper . Kao et al. proposed using relevance of the subject, reputation of the users, authority in category by building knowledge profile [12]. Shah et al. emphasized on the quality of answers to the questions and developed a prediction model for the best answers of the users [13].Pal et al. in their work presented a generalized definition of the experts as those providing a number of quality answers on a CQA forum. They used probabilistic and machine learning approach in order to find experts and those having the potential to be one[14]. Movshovitz et al. provided a reputation system to make experts attached to the website [15]. Based on the question and answers of the users as an attribute they used the random forest classifier to classify the experts from the non experts. The work also helps to predict the long term contributors on such a CQA forum.

Although most of these works consider the popularity of the questions and responses provided by the users, none of these techniques attempt to evaluate the difficulty level of the questions that are being responded by the users in determining the expertise. We believe that satisfactorily answering difficult questions along with the popularity of the questions and responses posted by a user can be a more suitable indicator of topical experts. In this paper, we aim to provide empirical evidence in support of this hypothesis.

III. PROPOSED APPROACH

In this section we describe our proposed approach of identifying experts in the StackOverflow community. Our approach attempts to solve the problems using generalized features of Difficult Questions (DQ) & experts. We propose certain features of DQ & experts and select the relevant ones, although the weightage of each feature is determined by the classification technique used. We use a random forest classifier for the same and compare the accuracy of the results for four different basis of expert labeling that we highlight in the later section. Before we proceed, we describe certain terminologies that we have used in the rest of the paper:

Z score [3]: Measures the answering tendency of the user, a and q are the number of answers and questions posted, respectively, by the user (In our case we take only topical posts).

z score = pa+q

Document Model for Textual Features Ranking [16]: It is a baseline method for ranking users based on their textual relevance to the documents(Topic & user profile). The probability that document d specifies that user e is an expert on query q is given as

p(qjd;e) = YP(tjd;e)n(t;q);

t2q

where t is a term in d and n(t;q) is the frequency of occurence of t in q.

Topic profile : The collection of all the keywords of the topics questions and answers.

User profile : The collection of all the keywords in the answers to the topical question of the user.

We would identify DQ first with appropriate features and include the number of difficult questions responded as a feature in our proposed expert detection technique.

A. Labeling Difficult Questions

We have defined two different aspects of the difficult questions in the Introduction section. To identify difficult questions whose solutions are not easily found, we propose a generalized set of features that we denote as FDQ. The feature list is detailed in table I. To identify difficult questions that is relevant to a large set of users in the CQA forum, we use the view count of the question as a relevant feature. A question whose view count is very high tends to be more difficult question than that having a low view count in the sense that former is fundamental and relevant to a large number of people. Thus we use view count as one of the distinguishing features of DQ. In table I, we provide an overview of some of these features to help understand their contribution in finding a DQ. We later observe that although none of these features alone may distinguish a DQ however a combination of one or more of these features uniquely identifies a DQ. As per the FDQ, for a question with a given value of ’Question Score’ or ’Comment Number’ the question may be difficult if the respective values are more than the standard deviation of respective group. Similarly ’Answer count’ for a question can indicate the difficulty level of the question if the answer count value on the question is less than the standard deviation value.

Sr.no	Features of Difficult Questions (FDQ)	Description
1	Question Score	Latest score earned by a question (M)
2	First response time	Time, duration between questions post time to its first answer’s time (M)
3	Comment Number	Number of comments on the post (M)
4	Answer length ratio	Ratio of the number of responses with term count more than the standard deviation of all responses of the question and the total number of responses
5	Questioner Z-score	Z-score of the questioner
6	Answerers Z-score ratio	Ratio of the number of responses with Z score more than the standard deviation of all responses of a question abd the total number of responses
7	Answer Count	Number of answers to the questions
8	Textual Rank	Document model used find the relevance of the questions based on the topical’s terms
9	Responder Z-score ratio	Ratio of the number of responders with Z-score more than the standard deviation of all responders z score of a question and the total number of responses
(M):Indicates modified as, where taken value = actual value - standard deviation value of the whole group’s data.		

TABLE I: Features of Difficult Questions

S.No	Features of Exerts (FEX)	Description
1	Questions score	score earned on all questions of,the user
2	Question count	count of All questions posted by the user.
3	Answer score	Score earned on all answers by the user.

4	Answer count	count of the answers posted by the uses on the same topic of questions.
5	User Z score	Z score of the user.
6	Question answer ratio	Question count / answer count.
7	Question Answer ratio with comments	count of questiont with (+)ve comments on them /count of all answers (+)ve comments.
8	Average post rate	Total score earned by user on its posts / total number of posts of user
9	Post rate	Rate of answering by the users i.e. total answers / time from joining till posting of last answer by the user.
10	Textual rank	Rank of user by document model(M).
11	Difficult questions answered	Number of difficult questions answered (as classified in the first case).
(M):Indicates modified as, where taken value = actual value - standard deviation value of the whole group's data.		

TABLE II: Features of Experts

of the whole group. Usually, questions with low responses are considered difficult ones. Thus, we use view count as one of the distinguishing features of DQ. For a difficult question, the First Response Time would be relatively higher than non difficult question. We consider 'Answer Length Ratio' as one of the features of DQ. We calculate the length of each answer in terms of keywords then find the standard deviation of the length values. This feature is actually the ratio of answer count having more number of words (length) than the standard deviation value of the whole group to the total number of the answers on the question. This feature gives a value more than 0.5 to DQ assuming difficult questions usually have explained answers. Similarly, we calculate 'Answerers Z Score Ratio' & 'Commentator Z Score Ratio' for each question. A questioner with high Z score tends to post a difficult question as it implies that although the user usually answers a lot of questions, however he has posted a question whose answer is not in his knowledge base. We define 'Textual Rank' of a question using document model of expert retrieval methods. This is explained above.

B. Features of an Expert

An expert is a user who has acquired a level of expertise on a particular topic. An expert may or may not answer an easy question but a difficult question is preferably answered by an expert. We present four different aspects of an expert.

Users who provides relatively large number of 'best answers'. This is the conventional definition of an expert in a CQA forum used in many previous works. For the stackoverflow.com 'best answer' is defined as the answer that a questioner selects out of many answers and credit answerer by tick marking that answer. This way a questioner gives his confirmation of meeting a solution to the problem he posted.

Users who provide answers to difficult question.

Users who fulfills the first criteria but also provides answer to the question where answer count is relatively low.

Users who fulfills first and second criteria both.

We denote the features of an expert as FEX that are highlighted in table II. Although, we might consider a user as an expert if the score of the user's post (questions or answers) is high. However, this is not always true, as experts might tend to answer questions that are raised by other topical experts who few in number are very and hence the scores obtained for such responses will be very low as compared to responses in popular but simple questions. Further, the count of user's questions and answers posted may also symbolize the expertise of the user.

As explained earlier, the Z score can be an identifying feature of the experts. We also calculate the Average Post Score of the user as an identifying feature of an expert. Higher the value of the average post score, the more likely he is to be an expert. The answering rate of an expert is also likely to be very high. The expert has a good 'Textual Rank' calculated using the topic and the user profile. Further, we have used the ratio of the number of responders with Z score above the standard deviation among all responders for a question and the total number of responders as an identifying feature. Finally, for an expert we use the number of DQs answered by him as a feature of an expert. In the next section we highlight the details of our experiments and the analysis of the results.

IV. EXPERIMENTS AND RESULTS

In this section we provide an overview of the StackOverflow data that we have used for our analysis and highlight the experimental procedure and results obtained. We obtain the StackOverflow data from www.data.stackexchange.com collected over a period of 2008/01/01 to 2014/01/01. For our analysis we have considered posts on three programming language-based topics, namely, c, java and python. Every data belongs to these topics considered and a complete topical isolation is maintained to get the true topical value for experts and DQ. The data consisted of 150,000 questions, approximately 500,000 answers and 210,000 comments from 50,000 questioners and 172,000 responders. We collected and modified the data for all features value listed in FDQ & FEX in tables I and II respectively. We have used random forest classifier (RF) with 10-fold cross validation technique for the classification over the identified features.

A. DQ labeling

We use a Random Forest classifier to classify the DQs based on the identifying features. We generate a labeled DQ data based on the view counts of the questions and represent the data set as List DQ 1. List DQ 1 contains each question under a common topic with its features value and a label of difficult or non-difficult question based on view count of the question when view count is more than the standard deviation of the group data. We prepare the training set by taking 10% of the entries in List DQ 1 and rest of the entries without label to prepare test set. We used RF to prepare the List DQ 2 based on best accuracy of classification model. List DQ2 contains RF classified entries. We present the result of comparison of List DQ 1 and List DQ 2 in table III and present the importance of the features in figure 1. Figures 1a and 1b shows the accuracy and F1 scores, respectively, when each of the features are taken independently and when they are combined.

B. Expert Identification

Similarly to label the experts we use four different basis as highlighted in the Introduction section and denote the labeled data set as List EX 1. List EX 1 contains each user under a common topic with its features value and a label of expert or non-expert based on different expert basis. We prepare a training set by taking 10% of the entries in List EX 1 and rest entries without label to prepare test set . We again use the RF classifier to prepare the List EX 2 based on best accuracy of classification model. List EX 2 contains RF classified entries. The result of comparison of List Ex 1 and List EX 2 in shown in Table IV. The results of the accuracy and the F1 score is shown in figures 2a and 2b respectively, when certain identifying features of the experts are considered independently and when they are combined. We next highlight the results obtained from our experiments and discuss the results.

C. Analysis and Discussion

We first state the results obtained from the RF classifier in identifying difficult questions.

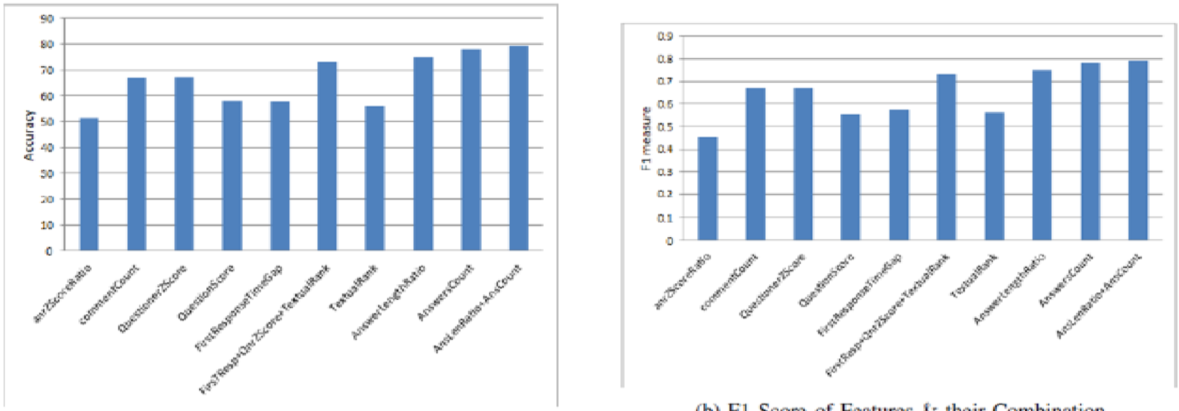
1) Identifying DQs: The classification of DQ shows an accuracy of above 95% with F1 measure of 0.9. This validates the authenticity of our classification method. It supports view count as a basis of measuring the difficulty of the questions. Upon checking manually we find that each question labeled DQ possesses one or more of the features FDQ. We observed that some of difficult questions are inter topical questions i.e. a question in a topic becomes a bit difficult in the topic when it involves the knowledge of other topics too. Better results are obtained when we use view count along with favorite count together as an assumption of DQ. The bar graphs in figures 1a and 1b, respectively, shows the accuracy and F1 measure of the features of DQ and their combination. Almost all accuracy values are above 50% indicating that the features are correctly selected for the DQ identification, however combining all the features provides the best accuracy.

2) Identifying experts: Table IV represents the result of the comparison of the List Ex 1 & List Ex 2. We present the results of four different aspects of an expert, the first basis being the more conventional approach of expert identification. We observe that the 3rd & 4th assumption of an expert out weights the conventional one. We observe that basis 3, where we combine ’Best Answer’ count with answer count on questions with relatively low answers, provides a higher accuracy and f1 measure with a reduced set of features as compared to the best answer basis (basis 1). However including the Difficult Questions feature (FEX 11) combined with the users ’Best Answers’ as criteria of expertise, provides the best results in terms of accuracy and F1 score. Thus this basis indicates a more reliable classifying model with much reduced set.

Thus using our proposed method we found better result than Balog et al. [16] that used the best answer feature. Dijk et al. [3] showed that Z score was not an important feature, however, in our experiments we find that Z-score plays an important role in expert detection depending upon the assumptions. However, accuracy & F1 score of basis 2 are relatively much lower indicating that only answering difficult questions may not be a good indicator for topical experts. Figures 2a and 2b measure of the accuracy & F1 score of the features of an expert when taken alone or in combination. We observe similar trends compared to the previous figures, where each

Topic & Difficulty Basis	Instances	Approx, matched instances %	Difficult questions correctly matched	% Difficult questions correctly matched	Accuracy	F1 measure
C , view count	45567	94 97	512 172	21.13 22.72	95.73 98.87	0.949 0.988
C, view count & Fav Count						
Python, view count	47374	95 97	415 119	22.92 19.32	97.57 99.31	0.974 0.993
Python, view count & Fav Count						
Java, view count	46877	93 96	529 216	21.63 21.79	93.40 98.769	0.961 0.987
Java, view count & Fav count						

TABLE III: Topicwise Result for Difficult Questions Classification over Two Different Basis of DQ



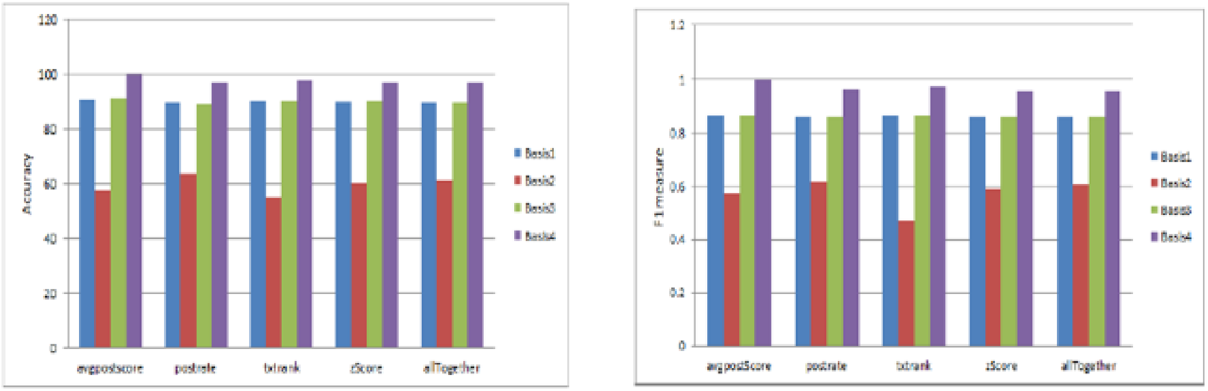
(a) Accuracy of Features & their Combination

(b) F1 Score of Features & their Combination

Fig. 1: Features contribution in Identification of DQ

Topic	Expert Basis	Instances	Approx % correctly matched instances	Accuracy	F1 measure
C	Best Answers	31681	99 49	94.772 54.791	0.926 0.547
	Difficult Questions		99	99.369	0.991
	Best Answers & Ans on Ques. Low Ans count				
	Best Answers & Difficult Questions		99	99.598	0.995
Python	Best Answers	33747	99 66	94.724 61.151	0.927 0.581
	Difficult Questions		99	97.536	0.967
	Best Answers & Ans on Ques. Low Ans count				
	Best Answers & Difficult Questions		99	99.321	0.99
Java	Best Answers	40140	97 64	82.75 54.981	0.773 0.544
	Difficult Questions		99	97.262	0.963
	Best Answers & Ans on Ques. Low Ans count				
	Best Answers & Difficult Questions		99	98.125	0.974

TABLE IV: Topic wise results of expert detection on four different expert basis. The table shows the result of four different aspects of the expert detection. We can easily observe that expert basis 3rd and 4th have out weighted the other two assumptions, considering that basis 1 has been the major feature of many works for detecting an experts.



(a) Accuracy of Features & their combination

(b) F1 Score of Features & their combination

Fig. 2: Features contribution in Detection of Experts

of the features considered independently has accuracy and F1 above 50% in both cases, however these values increases to nearly 99% when the eaftures are combined.

V. CONCLUSION

In our paper, we presented an approach to label difficult questions and using the capability of answering difficult questions for identifying topical experts in a CQA forum. We use a semi-supervised machine learning approach with a net gain of accuracy ranging from 5% to 16% over previous conventional approach. Although we have used the features provided in StackOverflow data however the features are generic and is present in most of the other popular CQA forums like Yahoo Answers and Quora. We established a relation between the generalized features of difficult questions and that of an expert. Other important features like temporality of the responses and knowledge gained over time needs to be considered for expert identification. However including these features would require different modeling approaches that we will subsequently look into as future goals of this work.

REFERENCES

[1] L. Mamykina, B. Manoim, M. Mittal, G. Hripcsak, and B. Hartmann, "Design lessons from the fastest q&#38;a site in the west," in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ser. CHI '11. New York, NY, USA: ACM, 2011, pp. 2857–2866. [Online]. Available: http://doi.acm.org /10.1145/1978942.1979366

[2] L. Yang, M. Qiu, S. Gottipati, F. Zhu, J. Jiang, H. Sun, and Z. Chen, "Cqarank: Jointly model topics and expertise in community question answering," in Proceedings of the 22Nd ACM International Conference on Information & Knowledge Management, ser. CIKM '13. New York, NY, USA: ACM, 2013, pp. 99–108. [Online]. Available: http://doi.acm.org/10.1145/2505515.2505720

[3] D. van Dijk, M. Tsagkias, and M. de Rijke, "Early detection of topical expertise in community question answering," in Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, ser. SIGIR '15. New York, NY, USA: ACM, 2015, pp. 995–998.

[4] B. V. Hanrahan, G. Convertino, and L. Nelson, "Modeling problem difficulty and expertise in stackoverflow," in Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work Companion, ser. CSCW '12. New York, NY, USA: ACM, 2012, pp. 91–94.

[5] G. Zhou, S. Lai, K. Liu, and J. Zhao, "Topic-sensitive probabilistic model for expert finding in question answer communities," in Proceedings of the 21st ACM International Conference on Information and Knowledge Management, ser. CIKM '12. New York, NY, USA: ACM, 2012, pp. 1662–1666. [Online]. Available: http://doi.acm.org/10.1145/ 2396761.2398493

[6] C. L. Lin, Y. L. Chen, and H. Y. Kao, “Question difficulty evaluation by knowledge gap analysis in question answer communities,” in Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on, Aug 2014, pp. 336–339.

[7] J. Liu, H. Shen, and L. Yu, “Question quality analysis and prediction in community question answering services with coupled mutual reinforcement,” IEEE Transactions on Services Computing, vol. PP, no. 99, pp. 1–1, 2015.

[8] F. Riahi, Z. Zolaktaf, M. Shafiei, and E. Milios, “Finding expert users in community question answering,” in Proceedings of the 21st International Conference on World Wide Web, ser. WWW ’12 Companion. New York, NY, USA: ACM, 2012, pp. 791–798.

[9] G. Cong, L. Wang, C.-Y. Lin, Y.-I. Song, and Y. Sun, “Finding question-answer pairs from online forums,” in Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ser. SIGIR ’08. New York, NY, USA: ACM, 2008, pp. 467–474.

[10] B. Wang, X. Wang, C. Sun, B. Liu, and L. Sun, “Modeling semantic relevance for question-answer pairs in web social communities,” in Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ser. ACL ’10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 1230–1238.

[11] G. A. Wang, J. Jiao, A. S. Abrahams, W. Fan, and Z. Zhang, “Expertrank: A topic-aware expert finding algorithm for online knowledge communities,” Decision Support Systems, vol. 54, no. 3, pp. 1442–1451, Feb. 2013.

[12] W.-C. Kao, D.-R. Liu, and S.-W. Wang, “Expert finding in question-answering websites: A novel hybrid approach,” in Proceedings of the 2010 ACM Symposium on Applied Computing, ser. SAC ’10. New York, NY, USA: ACM, 2010, pp. 867–871.

[13] C. Shah and J. Pomerantz, “Evaluating and predicting answer quality in community qa,” in Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, ser. SIGIR ’10. New York, NY, USA: ACM, 2010, pp. 411–418.

[14] A. Pal, F. M. Harper, and J. A. Konstan, “Exploring question selection bias to identify experts and potential experts in community question answering,” ACM Trans. Inf. Syst., vol. 30, no. 2, pp. 10:1–10:28, May 2012. [Online]. Available: http://doi.acm.org/10.1145/2180868. 2180872

[15] D. Movshovitz-Attias, Y. Movshovitz-Attias, P. Steenkiste, and C. Faloutsos, “Analysis of the reputation system and user contributions on a question answering website: Stackoverflow,” in Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ser. ASONAM ’13. New York, NY, USA: ACM, 2013, pp. 886–893.

[16] K. Balog, Y. Fang, M. de Rijke, P. Serdyukov, and L. Si, “Expertise retrieval,” Found. Trends Inf. Retr., vol. 6, no. 2&#8211;3, pp. 127–256, Feb. 2012. [Online]. Available: http://dx.doi.org /10.1561/ 1500000024