Exploiting User Activities for Answer Ranking in Q&A Forums

Chenyang Zhao(✉), Liutong Xu, and Hai Huang

School of Computer Science, Beijing University of Posts and Telecommunications, Beijing, China
chyzhao@bupt.edu.cn

Abstract. Many Q&A forums suffer high variance in the quality of their contents because of their loose edit control. To solve this problem, many methods are proposed to rank answers based on their quality. Most existing works in this domain focus on using variable features or employing machine learning techniques to automatically assess the quality of answers. Few of these works noticed that the relationship formed by user's activities can be helpful in capture the expertise of users in a specific topic. In this paper, we consider the relationship between users's activities in answer ranking task, create three new topic-aware features based on user profile information and the network formed by user's question-answering and comment activities, then we combine new created features with texture, user, comment features together and adopt a pairwise L2R approach SVMRank to rank answers. Experiments on a dataset extracted from Stack Overflow show that, (a) the new created features can better capture the expertise of users than other user features in answer ranking task. (b) the answer ranking approach get better performance when adding our new created features to the features used in previous works.

Keywords: Q&A forums · Answer ranking · Learning to rank

1 Introduction

Question and Answer Forums(Q&A Forums), such as Stack Overflow[1], Yahoo! Answers[2] and Quora[3] are characterized by loose edit control, which allows anyone to freely publish and edit almost everything, the varying quality of their contents has raised much concern.

To extract high quality contents from these forums, Q&A forums adopt voting mechanism, where users can indicate the quality of contents and even the reputation of the editors, or give the asker the right to select one of the answers as the best answer. While voting mechanism is effective in deemphasizing the low

[1] https://stackoverflow.com/.
[2] https://answers.yahoo.com/.
[3] https://www.quora.com/.

I. Romdhani et al. (Eds.): CollaborateCom 2017, LNICST 252, pp. 693–703, 2018.
https://doi.org/10.1007/978-3-030-00916-8_63

quality answers, it is dependent on the users's votes to the high quality answers. As a consequence, for many questions in Q&A forums, a high quality ranking of the answers is not provided. Our preliminary study on a dataset extracted from Stack Overflow shows that more than 60% of the questions without selected best answer, and more than 50% of questions the sum votes quantity of its answers is less than 200, which is not enough to get reliable ranking results.

Since such questions would largely benefit from ranking algorithms based on automated quality assessment strategies, in this paper, we propose a learning to rank approach (L2R) to rank answers in Q&A forums according to their quality. Unlike previous works, instead of directly estimating answer quality, we try to use the information of the answerer and the relationship between their activities to get better ranking results.

We first use the combination of up and down votes users got in a specific topic to create a new feature, then we combine the new feature with the question-answering network and comment network of user's activities to create other two features. At last, we combine the new created features with some other features we studied to get better ranking performance. Zhou *et al.* [7] found that almost all users are experts on only one or two topics, so all the methods to create new features and answer-ranking approaches in this paper are topic-aware. Experiments on a Stack Overflow dataset show that our new created features are significantly helpful in ranking answers in Q&A forums. Combining our new created features with features used in previous works, we get the best ranking performance.

2 Related Works

Answer ranking task attracts more and more attention in recent years along with the explosive growth with Q&A forums. Answer ranking is different from traditional Q&A system which is to generate an answer automatically, but to find a set of best answers among a list of answer candidates with various features.

Some researchers exploited a number of features to predict answer quality for ranking. Jeon *et al.* [3] built a framework to predict answer quality with non-textual features on maximum entropy approach and kernel density estimation. They also incorporated the quality scores into language modeling-based model and achieved significant improvements. Bian *et al.* [1] defined question-answer pair features to find high quality information in social media environment. They combined these features into Perceptron ranking model and achieved considerable improvements in accuracy.

User profile information attracts more and more attention as effective features in answer ranking task. Zhou *et al.* [7] extracted three groups of features from user profile information and tested their performance. Dalip *et al.* [2] summarized 97 features used in previous works and proposed 89 new features, devided them into eight groups, then applied them into Random Forest method to rank answers. They compared the performance of eight groups of features, found that user features is most helpful in answer ranking task.

Few of these works noticed that the relationship of user's activities can be helpful to capture the expertise of users in a specific topic. In this paper, we consider user profile information and the network formed by user's activities to create new topic-aware features and apply them into answer ranking task.

3 Create New Features

To create new features that can accurately represent user's expertise in a specific topic, we do an experiment to compare the performance of some commonly used user features in this domain, we use them individually as the basis to rank answers. The experiment results in Table 1 show that the quantity of up votes user's answers got in a specific topic is better than other information of the user in ranking answers. Besides the up votes, the down votes are also useful to capture the expertise of users. So we have Eq. (1) as the new feature and reffered it to as $F1$:

Table 1. The ranking performance of some user features. The meaning of features in the table are shown in Sect. 4.4. The meaning of metrics are shown in Sect. 5.2.

Feature	MRR	P@1	P@2
uf_{repu}	0.3487	0.2345	0.3976
uf_{uvac}	0.3767	0.2921	0.4293
uf_{ac}	0.3513	0.1932	0.3190
uf_{qc}	0.2830	0.1691	0.2883

In this equation $UpVotes(u_i)$ is the amount of up votes user u_i got and $DownVotes(u_i)$ is the amount of down votes user u_i got. All these values are topic-aware, which means they are calculated in a specific topic. Then we change the α to get the best ranking performance, it is shown that when α is 4.5, we get the best ranking accuracy.

$$F1(u_i) = UpVotes(u_i) - \alpha * DownVotes(u_i) \tag{1}$$

It also shows that the ranking performance when we only use the $F1$ feature is not good enough, so we combine it with the question-answering network and the comment network of user's activities to create another two features.

The question-answering network is a network that describes the relationship of user's question-answering activities. As shown in Fig. 1(a), we use $q_j \leftarrow u_i$ to represent u_i answering q_j.

With this network, a question can be seen as a hub of users and it is given a hub value to indicate its quality. And also we give every user an authority value to indicate his expertise. Then we have the following assumptions: (a) a good question will attract many expert users to answer it, (b) a expert user is the one

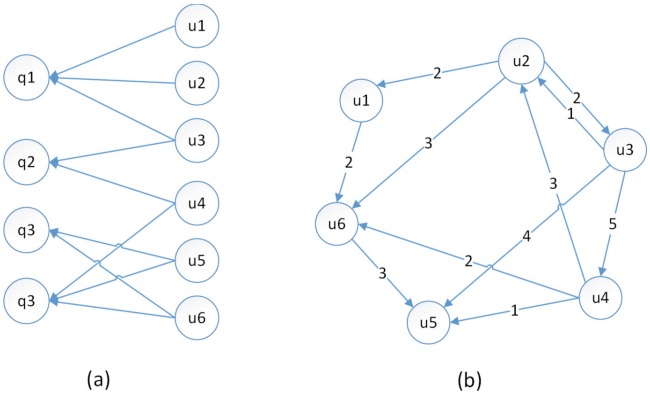

Fig. 1. The graphical representation of question-answering network and comment network.

whose answers are contained in many good questions. Then we use the HITS [5] algorithm to create a new user feature and refer it as to $F2$.

The other network we use in this work is the comment network between users. In Q&A forums, users can publish their comments to questions, answers and other comments. Because we only care the quality of answers, we only use the comments that posted to answers. As shown in Fig. 1(b), we use $u_i \rightarrow u_j$ to represent that user u_i commented the answer published by u_j and the integer in the arrow is the quantity of comments. The PageRank [6] algorithm is used to combine $F1$ with the comment network, then we get a new feature reffered to $F3$.

The three new created features are comprehensive combination of user's up votes, down votes and the relationship of their activities such as question-answering and comment, they are all topic-aware and they can well capture the expertise of users in a specific topic.

4 Other Features

In this section, we present other features used to represent the answer quality. These features try to capture the quality of the answer either directly, through texture features, as well as indirectly, through non-texture features, such as author profile information and comments to the answer. Totally, we study 43 features and organize them into groups according to the characteristics they try to capture. Thus, the features are devided into texture features, user features and comment features.

4.1 Texture Features

Texture features are one of the most successful and commonly used indicators of the answer quality in Q&A forums. Some of them are features about the

length of text, the general intuition behind them is that a mature and good quality text is probably neither too short, which could indicate an incomplete topic coverage, nor excessively long, which could indicate a verbose content. Some other texture features try to capture the structure of the answer text, they try to describe the answer quality directly through, analyzing the use of images, separation into sections, links and HTML format tags. A good answer should have the following attributes: (a) relevant to the question. (b) organized into sections, contains images and quoted blocks to improve understanding. (c) link to additional information for further study. Because the dataset we use is extracted from Stack Overflow, which is a Q&A forum serves for programmers, we use tf_csc to capture the use of code snippets in the answers. All texture features used in this work are listed in Table 2.

Table 2. Texture Features

Symbol	Description
tf_wc	n. of words in the text
tf_sc	n. of sentences in the text
tf_cc	n. of characters in the text
tf_avgpl	Average paragraph length
tf_cl	n. code lines
tf_avgcl	Average n. of code lines per code snippet
tf_pc	n. of paragraphs in the text
tf_ic	n. of images in the text
tf_lc	n. of links in the text
tf_outlc	n. of links to external sources
tf_inlc	n. of links to other questions/answers in the same forum
tf_lic	n. of lists in the text
tf_liic	n. of list items in the text
tf_qbc	n. of quoted blocks in the text
tf_stc	n. of $< strong >$ tags
tf_csc	n. of code snippets
tf_bm25	BM25 ranking function, based on a probabilistic retrieval framework
tf_ssc	n. of sentences shared by question and answer
tf_swc	n. of words shared by question and answer
tf_wciso	n. of words which appear in question and answer in the same order
tf_ldbw	Largest distance between two words that appear in answer and question

4.2 User Features

User features are frequently used in this domain in recent years and get very good performance in answer ranking task. The intuition behind user features is to indirectly infer the answer quality by examing the user who post it. More specifically, we are interested in features related to the user profile or its behavior, captured from events such as (a) posting questions and answers, (b) posting comments to questions and answers, (c) gain of merit votes and badges for questions and answers. In Table 3, we present all the user features computed for each answer.

4.3 Comment Features

Comment features try to capture the point of view of others to the answers, it is an important sign to show the answers are in high quality or not. All comment features used in this work are listed in Table 4.

Table 3. User features

Symbol	Description
uf_ac	n. of posted answers
uf_qc	n. of posted questions
uf_cac	n. of comments posted to answers
uf_cqc	n. of comments posted to questions
uf_uvac	n. of up votes to posted answers got from other users
uf_dvac	n. of down votes to posted answers got from other users
uf_avgcac	avg n. of comments per posted answer
uf_avgcqc	avg n. of comments per posted question
uf_avgac	avg n. of answers per posted question
$uf\_avguvac$	avg n. of up votes per posted answer
$uf\_avgdvac$	avg n. of down votes per posted answer
uf_bac	n. of posted answers selected as the best answer
uf_avgrp	avg rank position for posted answers
uf_tcq	n. of topics in which user post questions
uf_tca	n. of topics in which user post answers
uf_tcc	n. of topics in which user post comments

These three groups of features are used in previous works and get good performance in answer ranking task, so we build three baselines use the three different groups of features.

Table 4. Comment features

Symbol	Description
cf_cc	n. of comments posted to the answer
cf_maxccs	max n. of comments in one comment session
cf_uc	n. of users who commented the answer
cf_sumcl	sum of comment text length of all comments posted to the answer
cf_avgcl	avg length of comments posted to the answer
cf_cla	n. of comments posted by the answerer

5 Experiments

In this section, we first describe the dataset used in our experiments, then we introduce the metrics we use to represent the ranking performance, at last we explain our experiments in detail.

5.1 Dataset

The dataset used in our experiments consists of contents extracted from Stack Overflow, a Q&A forum for programmers. It consists of two parts, one part is about questions, answers and comments. Our methods to create new features and the approach to rank answers are topic-aware and in this work and we take the first tag of the question as its topic. We extract six different topics of questions and related answers and comments as our dataset. The detail of it is shown in Table 5.

The other part of the dataset is about users and their activities. To create the question-answering network and the comment network, we consider the Stack Overflow users who interacted with the users posted questions, answers and comments in the first part of the dataset. This part of dataset consists of 1356745 users, 13457828 question-answering pairs and 7112213 comments between these users.

Table 5. The detail of the dataset of questions, answers, comments and votes.

Statistics	Value					
	Topic-1	Topic-2	Topic-3	Topic-4	Topic-5	Topic-6
N. of questions	2045	2221	2079	2105	2160	1985
N. of answers	33569	38990	37653	35584	36535	30783
N. of comments	43420	50893	48540	44328	39893	36783
N. of upvotes	1229250	1233284	1185430	1256348	1283356	1055638
N. of down votes	84285	104782	82023	78876	103593	67328

As the same as some previous works, We use the ranking results based on quantity of up votes received by answers as the ground truth.

5.2 Metrics

Two standard information retrieval metrics are adopted in this work to evaluate the ranking performance as follows:

Precision@K(P@K): The K stands for the position of the correct answer. The precision at K reports the proportion of answers of the answer results set that has the corrent answer in position K.

Mean Reciprocal Rank (MRR): The reciprocal rank of a query response is the multiplicative inverse of the rank of the first correct answer. The mean reciprocal rank is the average of the reciprocal ranks of the results for a sample of queries. For a given query set Q, we calculate the MRR from below formula:

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

where $|Q|$ is the number of test questions and $rank_i$ is the position of the best answer in the ground truth.

5.3 Experiments and Results

To prove that the new created features can better represent the expertise of users in a specific topic, we first rank answers use the created features and other common used features like uf_ac, uf_uvac and uf_repu introduced in Sect. 4 as basis individually. Results are shown in Fig. 2.

The results show that the three new created features are all get better ranking results than other commonly used features in previous works, and the $F2$ feature get the best ranking performance in all the tested features. This indicates the effectiveness of the relationship of user's question-answering activities and comment activities in answer ranking task.

To compare the ranking performance of different groups of features described in Sect. 4 and the performance when the new created features are added, we adopt SVMRank [4] as the ranking algorithm to evaluate the performance of different combinations of feature groups. Table 6 shows all the combinations of features and experiment results. The SVMRank algorithm is also used in the work published by Zhou *et al.* [7] and achieved good results.

From the experiment results, we can see that in our three baselines user features get the best performance in ranking answers, that means user profile information is most helpful in answer ranking task. Most importantly, we get better ranking results when adding three new created features to any group of features used by previous works, for example, the accuracy is increased by 12.4% when $3F$ is added to user features. In particular, the combination of user features, structure features, length features, relevance features, comment

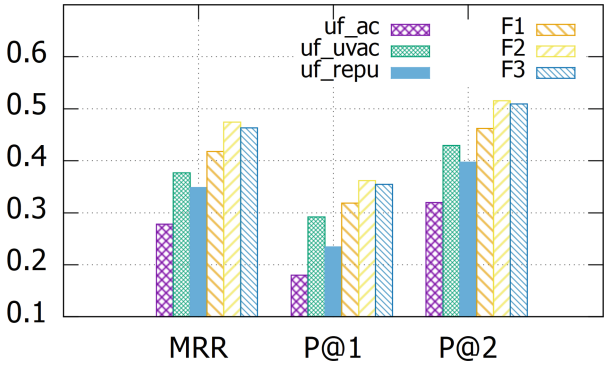

Fig. 2. The graphical representation of the performance of new created and user features when used individually in answer ranking task.

Table 6. The ranking performance of different combinations of feature groups. For simplicity, we use the combinations of capitalized first letter of every feature group to represent it, the three new created features are represented by $3F$ and $All = \{TF + UF + CF + 3F\}$.

No.	Comb.	Training Data						Metric.
		30%	40%	50%	60%	70%	80%	
1	TF	0.6521	0.663	0.675	0.6875	0.6983	0.7021	MRR
2	UF	0.6933	0.7021	0.7145	0.7188	0.722	0.7311	
3	CF	0.62	0.6399	0.6549	0.6675	0.6689	0.6731	
4	TF+3F	0.7156	0.7349	0.7499	0.7589	0.7691	0.7634	
5	CF+3F	0.6539	0.6678	0.6784	0.6875	0.6954	0.7022	
6	UF+3F	0.7265	0.7356	0.7523	0.7642	0.7751	0.7765	
7	TF+CF+3F	0.7431	0.759	0.7689	0.7742	0.779	0.785	
8	TF+UF+3F	0.7763	0.7869	0.798	0.8021	0.8093	0.8132	
9	**All**	**0.798**	**0.8005**	**0.8154**	**0.8245**	**0.8287**	**0.8396**	
1	TF	0.4231	0.4266	0.4336	0.4428	0.45	0.4579	P@1
2	UF	0.4758	0.4803	0.4889	0.496	0.49	0.5	
3	CF	0.3564	0.3598	0.365	0.369	0.3875	0.3986	
4	TF+3F	0.4657	0.4735	0.4806	0.5032	0.511	0.5198	
5	CF+3F	0.3975	0.4065	0.4135	0.4245	0.432	0.442	
6	UF+3F	0.5064	0.5138	0.5237	0.54	0.555	0.562	
7	TF+CF+3F	0.4981	0.5064	0.5123	0.519	0.5288	0.5372	
8	TF+UF+3F	0.5342	0.5451	0.5576	0.5673	0.569	0.573	
9	**All**	**0.5532**	**0.5669**	**0.578**	**0.582**	**0.592**	**0.6073**	

features and 3F get the best ranking performance, the accuracy of it is 21.4% higher than user features, which is the best of our three baselines. In general, the experiment results prove the effectiveness of three new created features, and show the helpful of user profile information and the relationship of user's question-answering activities and comment activities in answer ranking task in Q&A forums.

To further improve the effectiveness of the proposed features in this work, we compare our method with another answer ranking method proposed by [2]. In that work, the features are devided into eight groups and the Random Forest method is adopted to rank answers, we refer this method as $RF8$. The experiment results are shown in Fig. 3, they indicate that the method proposed in this work get better answer ranking performance, particularly, when 80% training dataset is used, the precision of our method is 6.2% higher than the method proposed by Dalip *et al.* [2] in answer ranking task in Q&A forums.

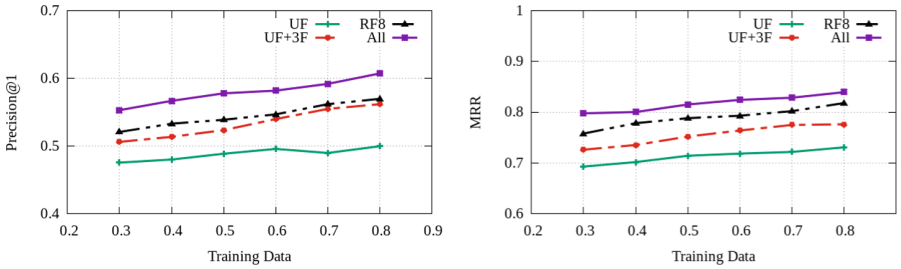

Fig. 3. The graphical representation of ranking performance of different combinations of feature groups and $RF8$.

6 Conclusion

In this paper, we create three new topic-aware features with user profile information and the relationship of user's activities to capture the expertise of users in a specific topic in answer ranking task and studied texture, user and comment features in answer ranking task. Then we compare the ranking performance of three groups of features and combine the new created features with these features to improve the ranking performance. The experiment results on a dataset extracted from Stack Overflow show that the new created features can better represent the expertise of Q&A users in a specific topic, and the new proposed method get better ranking performance than the state of the art method in answer ranking task in Q&A forums.

References

1. Bian, J., Liu, Y., Agichtein, E., Zha, H.: Finding the right facts in the crowd: factoid question answering over social media. In: Proceedings of the 17th International Conference on World Wide Web, pp. 467–476. ACM (2008)
2. Dalip, D.H., Gonçalves, M.A., Cristo, M., Calado, P.: Exploiting user feedback to learn to rank answers in q&a forums: a case study with stack overflow. In: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 543–552. ACM (2013)
3. Jeon, J., Croft, W.B., Lee, J.H., Park, S.: A framework to predict the quality of answers with non-textual features. In: Proceedings of the 29th Annual International ACM SIGIR conference on Research and Development in Information Retrieval, pp. 228–235. ACM (2006)
4. Joachims, T.: Optimizing search engines using click through data. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 133–142. ACM (2002)
5. Kleinberg, J.M.: Hubs, authorities, and communities. ACM Comput. Surv. (CSUR) **31**(4es), 5 (1999)
6. Page, L., Brin, S., Motwani, R., Winograd, T.: The page rank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab (1999)
7. Zhou, Z.M., Lan, M., Niu, Z.Y., Lu, Y.: Exploiting user profile information for answer ranking in cQA. In: Proceedings of the 21st International Conference on World Wide Web, pp. 767–774. ACM (2012)