Exploiting User Profile Information for Answer Ranking in cQA

Zhi-Min Zhou
Department of Computer Science and Technology
East China Normal University
Shanghai 200241, P.R.China
51091201052@ecnu.cn

Man Lan*
Department of Computer Science and Technology
East China Normal University
Shanghai 200241, P.R.China
mlan@cs.ecnu.edu.cn

Zheng-Yu Niu
Baidu, Inc.
Beijing, P.R.China
niuzhengyu@baidu.com

Yue Lu
Department of Computer Science and Technology
East China Normal University
Shanghai 200241, P.R.China
ylu@cs.ecnu.edu.cn

ABSTRACT

Answer ranking is very important for cQA services due to the high variance in the quality of answers. Most existing works in this area focus on using various features or employing machine learning techniques to address this problem. Only a few of them noticed and involved user profile information in this particular task. In this work, we assume the close relationship between user profile information and the quality of their answers under the ground truth that user information records the user behaviors and histories as a summary. Thus, we exploited the effectiveness of three categories of user profile information, i.e. engagement-related, authority-related and level-related, on answer ranking in cQA. Different from previous work, we only employed the information which is easy to extract without any limitations, such as user privacy. Experimental results on Yahoo! Answers manner questions showed that our system by using the user profile information achieved comparable or even better results over the state-of-the-art baseline system. Moreover, we found that the picture existence of a user in cQA community contributed more than other information in the answer ranking task.

Categories and Subject Descriptors

H.3.5 [**INFORMATION STORAGE AND RETRIEVAL**]: On-line Information Services—*Commercial services, Web-based services*

Keywords

Community-Question Answering, user profile information, answer ranking

1. INTRODUCTION

In recent years, *Question Answering* (QA) has significantly changed due to the bloom of community question-answering (cQA) sites, such as Yahoo! Answers[1], Baidu Zhidao[2] and Quora[3]. Many user-generated contents (UGC) without standard expressions and formats have been created in these communities by different people for various purposes. As a result, these UGCs accelerate the research of non-factoid question answering. Most previous work aimed only on factoid question answering, mainly regarding to name entity recognition and specific pattern recognition. However, to address non-factoid question answering, these simple recognitions are far more insufficient.

Previous work attempted the features from factoid question answering for non-factoid question answering and got promising performances with some textual and non-textual features[16]. However, these common off-the-shelf NLP features are too complicated to be easily calculated and even more sometimes they are inaccessible for all researchers, i.e. web query log.

On the other hand, almost all existing cQA sites develop many incentives, i.e. user scores or hierarchy, to increase their stickiness and keep their community fresh. As a result, both question askers and answer providers tend to stay for a long time and return to the community. In this situation, user information such as user behaviors in one session or user history are dramatically increased. Recently, some studies proved that the user information can be used to predict the answer quality in cQA [9, 15, 8, 1, 14]. This is crucial since the answer quality is an important prerequisite of answer ranking.

Our preliminary study on Yahoo! Answers showed that the number of best answer provider follows the Long Tail theory. That is, about 20% of users present almost half of the best answers. As a reward, these users would be assigned higher scores and higher best answer rate, which can be found in their user profile information. Another finding is

*Corresponding author.

WWW 2012 Companion, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1230-1/12/04.

[1]http://answers.yahoo.com
[2]http://zhidao.baidu.com
[3]http://www.quora.com

that most of those top contributors are always good at only one or two question categories. Although these findings are quite natural in our daily life, it is still interesting when we discover them in cQA online too.

Thus, inspired by previous work and these findings, we address the problem of answer ranking of non-factoid question for cQA using effective methods in factoid / non-factoid QA in combination with the user profile information in cQA. Specifically, the purpose of our work is to answer the following two questions:

1. Can user profile information help to improve answer ranking in a cQA service?
2. What kind of user profile information can be helpful in a cQA service?

Our overall approach is to analyze and explore the user profile information for answer ranking on Yahoo! Answers data. Specifically, for the first question, we first follow one previous work in [16] on Yahoo Webscope data set to build a baseline system and then we incorporate the user profile information extracted from the same data set to re-rank the results. For the second question, we extract new question-answering pairs in one specific week from Yahoo! Answers to explore the effectiveness of various user profile information for answer ranking. For clarity, we also group different user profile information into three categories, i.e. level-related, engagement-related and authority-related respectively, according to their sources. To the best of our knowledge, few research work involves user profile information, such as the level, picture existence, previous best answer collection etc., for answer ranking.

We performed evaluations on baseline systems and re-ranking systems on Yahoo! Answers manner questions in the first experiment. The experimental results showed that using user profile information can achieve a comparable or even better MRR and P@1 over the state-of-the-art features. Moreover, we also found that the user engagement-related information has the highest information gain followed by the user authority-related information for answer ranking task.

The rest of this paper is organized as follows. Section 2 introduces the empirical study of two main cQA services. Section 3 describes our methods. Section 4 presents our experiments and results. Section 5 reviews the related work. Section 6 concludes this work.

2. EMPIRICAL STUDY OF CQA SERVICE

2.1 Yahoo! Answers

Yahoo! Answers is probably the largest open-domain knowledge sharing community where people can ask or answer questions on the web. It has 27 main categories containing various topics, ranging from our daily life to sophisticated scientific knowledge. There are hundreds of millions of unique users worldwide to ask or answer questions on Yahoo! Answers. Some of them are experts who can provide high quality answers with their expertise, but most of them just share personal experience when looking for answers to their questions. To encourage more users participating in the community as well as providing high quality answers, Yahoo! Answers sets up an incentive system based on points and levels. Basically, one can get high levels or top ranks by continuously providing high quality answers. Although the point associated with one user cannot be used to buy or redeem anything, it does make other users to identify how active and helpful he/she is.

2.2 Quora

Different from Yahoo! Answers, Quora is more likely to be a collection of questions and answers organized by everyone who use it. In Quora, the question asker is not necessarily expected to be the best judger of the given answers. That is, Quora never rates an answer but encourage everyone to participate in revising the answers. Unlike the traditional cQA service, this design actually helps to get high quality answer candidates because the users of Quora are those people called as Elite, such as technique pundits, marketers, social media mavens. Moreover, they incline to provide high quality answers or suggestions by using their own authentic identities or real names.

2.3 Users in cQA services

Despite the difference between these cQA services, there is no doubt that users are the crucial parts of a cQA service besides questions and answers. Obviously, different cQA services provide distinct user experience and cultivate various user habits as well. For example, the behavior of a user in Yahoo! Answers can be described in a waterfall flow as follow.

$$Ask\ question -> \ Answer\ question -> $$
$$Discover\ best\ answer -> \ Close\ question$$

Thus they have no chance to modify previous answers or add a new answer if the question is resolved by people or system. While in other cQA services like Quora, people are likely to be more active in creating or editing the answers. For example, they can modify incomplete answers, remove an irrelevant answer from other users, or even refine the question to make it better. Since people prefer optimizing a fair answer rather than put it in the waiting list, it is more likely to get high quality and authoritative content from Quora. However, people would prefer raising an urgent question on Yahoo! Answer when they want to be anonymous, i.e. how to do an abortion.

Another difference between these cQA services is the way they instruct people knowledge. For example, Yahoo! Answers handles each question within a fixed response time. After that since the question cannot be answered by users any more, people may not get a complete answer timely. Nevertheless, Quora emphasizes the logic or completeness of the solution rather than the response time. Consequently, this difference affects people on learning unknown things and changes their engagement on cQA service. Figure 1 depicts the distinct characteristics of users of different cQA services.

3. OUR APPROACH FOR ANSWER RANKING

3.1 Extract User profile information

Almost all existing successful cQA services rely on millions of users and their active communication. No matter whether the users use real identity or not, their unique behaviors and statistics are recorded in user profiles. From previous work, we found that those user information can be used to

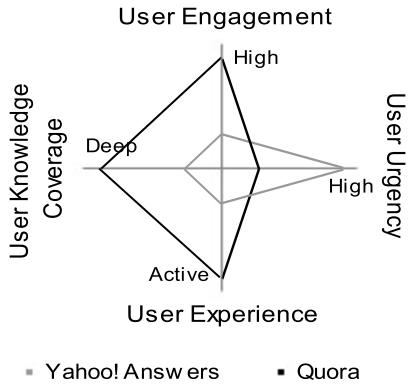

Figure 1: Distinct characteristics of users from Yahoo!Answers and Quora.

evaluate or predict answer quality and even the satisfaction of question asker. For example, [9] [8] and [15] explored several answerers characteristics, i.e. user level information such as total points, authoritative information such as best answer rate, on the evaluation of answer quality. Moreover, [10] exploited how user profile information can be used to predict the satisfaction of external web searcher.

Inspired by their work, we try to employ user profile information on answer ranking. Figure 2 is the screen shot of one specific user's profile web page. When we take a close look at the defined user profile page of Yahoo! Answers, we find that most of user information is explicitly presented on the web page, i.e. points, levels, number of answers answered, etc. These information records the histories and behaviors of one specific user. On the other hand, these basic information can also reveal additional implicit information, such as user engagement, popularity and authority, when they are combined together. That is, although these implicit information cannot be directly retrieved from user's profile web page, we can acquire them based on the basic explicit information. In consideration of above analysis, we build up an information system consisting of three categories to model the information of users. They are **level-related(LR)**, **engagement-related(ER)**, and **authority-related(AR)** information. We next present these three categories of information in details.

3.1.1 Level-related information

Level - Level information is derived from the points one user gets in Yahoo! Answers. The more points one user accumulates, the higher level and the less limitation such as comments, ratings, votes per day he gets. There are 7 levels ranging from 1 to 7.

Points - Points are collected from the users' actions on the community. For example, if one user logs in to the community in the first time today, he can get one point. However, each time one of the following two actions will expend one point: 1) ask a question or 2) delete an answer.

Total Answers - the number that one user has answered so far, i.e., the sum of best answers and other answers.

Answer-Question rate - the proportion of the number of total questions and the number of total answers.

3.1.2 Engagement-related information

Points earned in current week - the points according to the users' actions performed in current week.

Picture existence - This is a Boolean value to indicate whether the user has a avatar or not.

Average points of one week - the points one user may usually get in one week. We normalized the total points with the member existence. Here the member existence means how long the user account has been created in the community by now.

User engagement - There are two types engagement of one user. One is the short-period engagement, and the other is the long-period engagement. The short-period measurement comes from the *Points earned in current week*, while the long-period measurement is derived from *Average points of one week*. Based on our analysis from top contributors in the Yahoo! Answers leader board, we find that if *Points earned in current week* $>$ 0.5 * *Average points of one week*, the status of user engagement is more likely to be active. Meanwhile, the effect of burstiness of Individual answering activities, such as no points in the holiday week, should be decreased. Finally we get the combined user engagement by applying the below formula.

$$Engagement = 1/3 * Average\ points\ of\ one\ week + 2/3 * (Points\ earned\ in\ current\ week - Average\ points\ of\ one\ week) \quad (1)$$

3.1.3 Authority-related information

Best Answer rate - the percentage of the number of best answers and the number of total answers.

Top contributor or not - an honorable label in the community indicating how much helpful one user is in answering question. If one user is labelled as top contributor, it means that he is ranked at the top 10 in at least one question category.

Expertise category number - This value only exists when the user is a top contributor. The high category number always reveals one user's active engagement as well as board expertise.

ExpertiseRank - This rank is borrowed from PageRank. Considering the top contributors or experts always give the best answers. If they vote for another answer candidate, the answer candidate may have more chances to be a correct answer or even a best answer. Thus, the more votes got from those experts, the more likely the users have expertise.

Perplexity score on previous threads - This information is derived from the perplexity score of the current answer calculated by trained language model on previous threads. This is quite reasonable because for each user, his vocabularies, sentence patterns and expertise knowledge would not change dramatically in a short period of time. Therefore, we can use the perplexity score to evaluate the similarity between one question and the answer provider's expertise.

Perplexity score on previous best answer collection - This information is quite similar with the perplexity score on previous threads. However, the only difference is the training corpus. Since the manner of most questions aims to inquire the solution step by step, it is more likely different answers have similar structure. Moreover, in most cases these answers with clear structure are more likely to be best answers.

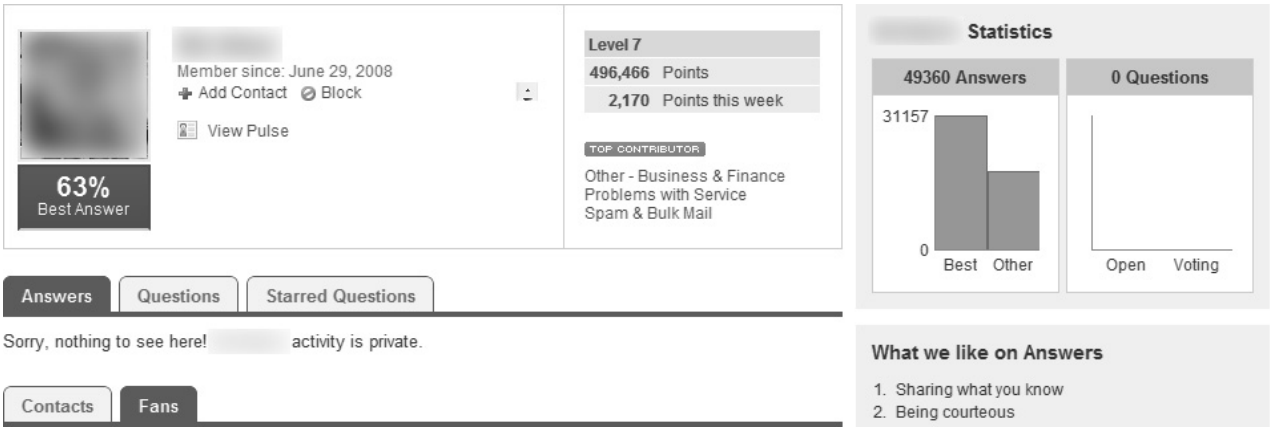

Figure 2: User profile sample

3.2 Ranking using user profile information as additional features

Previous work investigated the effectiveness of several state-of-the-art QA features for non-factoid answering ranking[16]. Due to the superior performance of these features in previous work and our data limitation, we adopted 6 types of features in our system to set up a benchmark. These 6 types of features are described as follows:

Similarity - Following previous work, we use the BM25 formula of Terrier IR platform[4] to extract this feature.

Translation - This feature indicates the probability that one question Q can be a translation of one answer A. We calculate this probability using GIZA++ toolkit[5].

Same word sequence - This feature counts the number of non-stop question terms matched in the same order in the whole answer.

Answer span - This feature counts the number of total words and noun words existing in the largest distance between two non-stop question terms in the answer.

Overall match - This feature records the number of non-stop question words appearing in the complete answer with no strict order.

Informativeness - This feature indicates the number of non-stop words in the answers. Usually, they are nouns, verbs and adjectives and should not be question terms.

To address the first question, we incorporate an authority-related information, ***Perplexity score on previous best answer collection***, as an additional feature in the baseline. The reason that we use this score only instead of the entire information is its stabilization. Unlike other user information, this score can naturally decrease the side-effect by many burstiness, i.e. user low engagement in one particular week. Moreover, this score also reflects the hidden pattern from the vocabulary and structure that people used.

Basically, different categories of questions and their answers have varied vocabularies and structures. For example,

E1) Q: *How do I travel from Chicago to Wisconsin dells?* Best Answer: *According to their site, Greyhound has one bus a day.*

E2) Q: *How to travel from Nasik to Bhimasankar?* BA: *227 kilometers, u can look for bus or car from nasik.*

E3) Q: *How do I get my dogs to stop barking?* BA: *give them a bone with meat on it.*

E4) Q: *How do I keep my Chihuahua's ears clean?* BA: *After her bath, use a few cotton balls to dry out her ears. Have the vet look at her ears, she may need a little medication/ear cleaner.*

The words, i.e. *travel*, *bus*, *from*, *to*, *cars* in E1 and E2, are typical words in Travel category. While the words, i.e. *vet*, *bone*, *dog* in E3 and E4 definitely indicate that the question-answer pair belongs to Pets category. It exhibits a very low degree of ambiguity of category. So we train different language models for each category to get rid of the noise from other categories. Then, we calculate the perplexity score of answer candidates using corresponding language model. Finally, we use this perplexity score as an additional feature after normalization.

3.3 Effectiveness of user profile information for answer ranking

Previous work showed the effectiveness of user information on question recommendation task, especially on finding experts in the community [18]. Meanwhile, experts or even users with some sort of expertise usually provide high quality answers than others. Considering the user profile information reflects all user behaviors and history statistics, we assume this information would be helpful for distinguishing the quality of answers as well as answer ranking task.

Following our user profile information system built for each user, we extract all three categories information i.e. level-related, engagement-related and authority-related information, from the users in our data set. Then we exploit the performance of different combinations of these features on answer ranking task. Since Yahoo! Answers protects user information due to privacy agreement, we can only access the following information to perform the experiment and investigate their effectiveness on answer ranking task: ***level***, ***points***, ***average points of one week***, ***picture existence***, ***points earn this week***, ***total answers***, ***best answer rate***, ***answer-question rate***, ***top contributor or not***, ***expertise category number***, ***user engagement***.

[4]http://ir.dcs.gla.ac.uk/terrier
[5]http://www.fjoch.com/GIZA++.html

# of categories have best answers	# of unique users	Percentage
1	13216	88.04%
2	1297	8.64%
3	307	2.05%
4	121	0.81%
5	45	0.30%
6	12	0.08%
7	6	0.04%
8	4	0.03%
9	2	0.01%
11	1	0.01%

Table 1: Percentage of unique users who provide at least one best answer in DS1

Statistics name	Value
# of unique user	1405
# of exists avatar	286
# of is top contributor	86
Mean of average of points one week	141
Points earn this week	111

Table 2: Overall profile information of the user in DS2

4. EXPERIMENTS AND RESULTS

4.1 Data sets

We use the Yahoo! Answers Manner Question, version 2.0[6] as our data set, denoted as *DS1*, to perform the first experiment. We fist select these questions beginning with *how (to|do|did| does|can|would|could|should)*. After filtering the low quality question-answer pairs, the final subsets contains around 150K questions and corresponding answers. Note the data set also remains some meta data, such as the category of the question and the best answer flag. Then we split the data set into training, development and test, containing 60%, 20% and 20% of questions respectively.

In the second experiment, we do not use *DS1* since it contains no user profile information due to the user privacy terms. Thus, we randomly crawled manner questions and their related user profile information of all categories from Yahoo! Answers to construct a new data set, denoted as *DS2*. For completeness, We also adopt the same preprocessing methods as the first experiment. Besides, we only select the questions which have been resolved within one specific week in order to keep all meta data of one user, for example, points of the current week, are recorded in the same time period.

Table 1 shows the statistics of the two data sets and lists the percentage of unique users who provide at least one best answer in *DS1*. From the table, we can see that almost 88% best answer providers in the data set give best answers in one specific category. And only around 1% of the best answer providers are good at 4 or more than 4 categories. Also, we find that around 8% of unique users have been rated at least 5 best answers in one category. That means most people have expertise only in one category. Table 2 summarized the overall profile information of the user in DS2.

[6]http://webscope.sandbox.yahoo.com/catalog.php?datatype=l

4.2 Preprocessing

To further improve the data quality for experiment, we first filtered the question-answer pair containing less than 10 answer candidates or more than 20 answer candidates since most people usually only read the first two pages (normally 20 records) of the results. Besides, question-answer pairs with 10-20 answer candidates appear more often than other number of pairs in the data set. After that we cleaned the data by removing user doodle and HTML tags. Finally we perform POS tagger on each answer candidate.

4.3 Evaluation metrics

We adapt the standard information retrieval metrics to evaluate the performance as follows.

- **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 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 queries, $rank_i$ is the position of the correct answer.

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

4.4 Ranking incorporate user profile information as additional features

We extracted the 6 features with high performance mentioned in 3.2 and followed extracting method in previous work[16] to build up a baseline system. Then, we used SRILM toolkit[7] to calculate one AR user information, *Perplexity score on previous best answer collection*. As we mentioned before, in order to decrease the noisy between different categories, we only train the language models based on lower-cased best answers in one specific category on the development set. After training 26 5-gram language models, we converted all words into lower cases and used the language model to calculate the perplexity on an entire answer candidate. Finally, we combined normalized perplexity score with the 6 features and their normalization features in our experiment.

We adopted two common ranking models, SVMRank and ListNet, to evaluate the performance of AR user information. Also we tuned the weight of AR information on the development set, and used the best combination as feature set on the test set.

Table 3 summarizes the results of baseline using state-of-the-art features and AR information on both development and test set. From this table, we can find our baselines using SVMRank and ListNet on the test set achieved comparable results with previous work[16]. The high MRR score shows the correct answer is often retrieved within first two answer candidates, and P@1+2 score tells us that over 90% of instances have been re-ranked successfully.

[7]http://www.speech.sri.com/projects/srilm/

	t	p-value
Pair: SVMRank@6F - SVMRank@AR	0.2731	0.7848
Pair: ListNet@6F - ListNet@AR	-4.3106	1.668e-05

Table 4: Significant difference between the results of baseline system and system using AR information only on two ranking models

On the development set, we can find the systems incorporated with AR information achieved the highest performances among all. For example, all two ranking models with AR information yield almost 1% improvement in terms of MRR score over the baseline systems. However, there is no much difference between weighted AR information and non-weighted AR information on SVMRank model. Unfortunately, the MRR and P@1 scores have no changes with the weight of AR feature on ListNet model. Another unexpected result comes from the high performance of the system using AR information only. On the SVMRank model, the systems using AR information only achieved the second highest both on development and test set. And the simple AR information performed comparable scores on MRR in both ranking models with baseline. Note that the AR information we used here only contains a small proportion of user information, this feature is quite promising.

Moreover, we performed a paired t-test between the results of baseline system and our system using AR information only (Table 4). Results show that there is no significant difference on SVMRank model between the two systems, which implies that the simple AR information can replace the complicated state-of-the-art features without decreasing the performance. On the other hand, the significant difference between two systems on ListNet model indicates that the bottleneck of the AR information itself. Since the baseline system performed better than the system using AR information only, there is no surprise that the big difference between them. As a whole, our results achieved a slight improvement over the state-of-the-art baseline system. All these results indicate that using user profile information can help to improve the answer ranking performance.

4.5 Effectiveness of user profile information for answer ranking

We build the system on the crawled data set DS2 and employ the different combinations of user information as features. For clarity and completeness, we use ER to represent the combination of *points earned in current week, picture existence, average points of one week* and *user engagement*, LR to represent the combination of *level, points, total answers* and *answer-question rate*, AR to represent the combination of *best answer rate, top contributor or not* and *expertise category number*.

Table 5 shows the performance of different combinations of three information categories. Generally, all the combination with at least two feature categories achieved MRR score above 0.4. However, the combination of all the information only obtains 0.412 in MRR, which is 0.5 lower than the best result. Also, the P@1 scores of these three combinations rank at the second place in experiment. The highest MRR of our system comes from the combination of two engagement-related user information, for example, {*points earned in current week*, *picture existence*}, or {*average points of one week*, *picture existence*}, or {*user engagement*, *picture existence*}. This finding is quite interesting and reasonable since we all know that the more the users care about the website, the more engagement they would involve. Therefore, the engagement-related user information, especially the *picture existence* information can provide important message to judge the quality of answer.

Comb.	Feature set	MRR	P@1	P@2
Baseline	6F	0.575	18%	67%
ER_2C	ptwk+pic	0.516	22%	36%
ER_2B	avgpt+pic	0.516	22%	36%
ER_2A	eng+pic	0.516	22%	36%
AR+LR	all	0.468	2%	88%
AR	all	0.468	2%	88%
LR+ER	all	0.412	26%	21%
ER+AR	all	0.412	26%	21%
ER_3B	eng+ptwk+pic	0.412	26%	21%
ER_3A	avgpt+ptwk+pic	0.412	26%	21%
ER	all	0.412	26%	21%
AR+LR+ER	all	0.412	26%	21%
LR	all	0.362	18%	12%
ER_3C	eng+avgpt+ptwk	0.244	6%	11%
ER_2F	avgpt+eng	0.244	6%	11%
ER_2D	ptwk+avgpt	0.244	6%	11%
ER_2E	ptwk+eng	0.231	6%	9%

Table 5: Results of combination information used in answer ranking on DS2. ptwk stands for points earn this week. pic stands for picture existence. avgpt stands for average points of one week. eng stands for user engagement.

Besides the high contribution by the engagement-related information, authority-related information also offers reasonable results compared with level-related information. Specifically, the P@1+2 scores achieved by two AR involved combinations are much higher than the score of baseline system. Considering the AR information is directly derived from users' expertise, it is not surprising that AR information can help in answer ranking task.

We also note that level-based information does not perform well in the experiment. One possible reason may due to the mechanism of points system in Yahoo! Answers. Firstly, the distribution of level information is quite uneven since there are only 7 values of this information. So it is impossible to distinguish users from inferior level to superior level despite of the huge difference between their engagements. Secondly, the level information is derived from the points information, which may not reflect the user behavior rightly. For example, the points system of Yahoo! Answers not only adds point for answering question, but also deducts point for raising a new question. As a result, if one user answers one question and then raises another new question, the system would deduct 2 points from his total points. Therefore, the more questions the user asks, the less points he will get. That is also the reason why we considered to involve answer-question rate and the number of total answers in the level-related information category. However, these two information are still too weak to improve the result.

In addition, we find the rest of the engagement-related information, i.e. *average points of one week, points earned in current week* and *user engagement* achieved the worst per-

	SVMRank				ListNet			
	Feature set	**MRR**	**P@1**	**P@2**	**Feature set**	**MRR**	**P@1**	**P@2**
Dev	6F+2500*AR	**0.5835**	19.97%	78.60%	6F+AR	**0.6139**	26.07%	70.10%
	AR	0.5811	19.68%	78.23%	6F+10*AR	0.6139	26.07%	70.10%
	6F+1000*AR	0.5798	19.63%	78.35%	6F+100*AR	0.6139	26.07%	70.10%
	6F+AR	0.5792	19.63%	78.38%	6F+1000*AR	0.6139	26.07%	70.10%
	6F (Baseline)	0.5787	19.59%	78.41%	6F+2500*AR	0.6139	26.07%	70.10%
	6F+100*AR	0.5726	19.59%	78.49%	6F (Baseline)	0.6091	24.91%	70.12%
	6F+10*AR	0.5709	19.22%	78.51%	AR	0.5872	20.01%	78.32%
	Systems	**MRR**	**P@1**	**P@2**	**Systems**	**MRR**	**P@1**	**P@2**
Test	Best of Ours	**0.5513**	10.72%	88.60%	Best of Ours	**0.5931**	20.96%	73.20%
	Only AR feature	0.5510	10.67%	88.32%	Baseline system	0.5891	20.04%	74.22%
	Baseline system	0.5506	10.59%	88.65%	Only AR feature	0.5510	10.77%	88.57%

Table 3: Summary of the performances on DS1 development and test data set. 6F stands for the features mentioned in 3.2

formance in both MRR and P@1,2. Since these information derived from the level-related information points, it can be explained by above analysis.

Furthermore, we performed an attribute selection among these information under the information gain evaluator on Weka[8]. Results shows that the three highest contributors are *picture existence*, followed by *best answer rate* and *top contributor or not*.

4.6 Analysis

Experimental results on Yahoo! Answers manner questions showed that the user profile information, especially the engagement-related(ER) and authority-related(AR) information significantly improve the performance of answer ranking. Our first experiment proved the effectiveness of the AR user information by achieving comparable results with state-of-art baseline system. Specifically, ranking only with the AR user information can achieved 0.0004 in MRR and 0.1% in P@1 improvements respectively over the baseline. Meanwhile, we achieved more than 0.1% improvement by incorporating the AR information with traditional NLP features over the baseline. However, in terms of MRR score, the AR information only achieve no more than 0.5510 regardless of the models used.

On the other hand, the second experiment shows that the picture existence information can further distinguish the best answer among the answer candidates. Specifically, the best results of our system achieved 0.06 lower in MRR than state-of-the-art baseline system. That means we can simply distinguish the quality of an answer candidate by observing just several information in the user profile as the first step. However, we should state that both the ER information and the AR information prefer users with frequent engagements rather than rookies. Thus, textual features should be also involved to balance the results.

5. RELATED WORK

Community answer ranking is different from traditional QA system which is to generate an answer automatically, but to find a best answer among a list of answer candidates with various features.

Some researchers exploited a number of features to predict the answer quality for ranking. For example, Jeon et al. [4] 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. Shah and Pomerantz[15] assessed the answer quality with 13 factors, proved contextual features and social information can help the prediction of answer quality. Liu et al. [10] proposed 37 features including asker satisfaction features to predict the answer quality, and they also made huge progress on question quality and user's satisfaction evaluation. Surdeanu et al. [16] made great efforts on investigating a wide range type of features, varying from similarity, translation, density and frequency to web correlation features. They combined these features into Perceptron ranking model and achieved considerable improvements in accuracy. A disadvantage of this model is lack of user information. Blooma et al.[2] proposed a light framework combined with both textual and non-textual features. The results suggested that the quality of the best answer was influenced by the textual features rather than non-textual features. Quarteroni [14] built a User Model representing individual users' reading level, age range and interests in answer filtering.

Other researchers tried this task by machine learning methods. Ko et al. [6] set up a unified probabilistic answer selection framework for question answering. They chose the best answer with the probability of relevance between an answer candidate and supporting evidence. Later, they improved their system by taking the joint probability of all correct answers into account. Wang et al. [17] involved analogical reasoning to reduce the lexical gap between questions and answers. They also defined different links from a Bayesian logistic regression model to detect a high-quality answer. Liu and Agichtein [11] used standard machine learning techniques to predict best answer from the information seekers' perspective. They assume asker's prior experiences, expectations and personal preferences can determine the latest selection of best answer. And results showed a personalized prediction model performs better over "one-size-fits-all" model. Moschitti and Quarteroni [12] trained with Support

[8]http://www.cs.waikato.ac.nz/ml/weka/

Vector Machines and string kernels, syntactic and shallow semantic tree kernels to perform answer selection.

On the other hand, some researchers proposed to rank the answerers to find the experts first, and then directly treated experts' answer as best answers. Liu et al. [8] proposed a mixed framework of language model and latent dirichlet allocation model with answerers' answering history and interest to find the appropriate best answerer to provide high quality answer. Jurczyk et al.[5] and Zhang et al. [18] tested a set of network-based ranking algorithms, including PageRank and HITS to identify experts and professional in the community. Nie et al. [13] also tried PageRank for authority investigation. Li et al. [7] proved that category information of question is crucial for question answering area, especially for question routing. They used category-sensitive language model combined with answerers' expertise to achieve a higher accuracies of routing question with lower computational costs relative to other state-of-art methods. Bouguessa et al. [3] proposed an automatically way to discriminate between authoritative and non-authoritative users.

Compared to existing works, we investigated and analyzed the effectiveness of user profile information for answer ranking. Moreover, our methods can both used to both labeled and unlabeled data.

6. CONCLUSIONS AND FUTURE WORK

In this paper we have exploited user profile information for cQA answer ranking. On one hand, we proved the user profile information extracted from Yahoo! Answers is helpful to evaluate the quality of answer candidates. On the other hand, we further investigated the effectiveness of various use profile information.

Different from previous work used many complex features, we tried the user profile information easily to be extracted without any limitation, such as user privacy. Results on Yahoo! Answers manner questions showed that using the user profile information, especially the authority-related information, achieved comparable results over state-of-the-art baseline system. Also we achieved 0.001 and 0.1% improvement in MRR and P@1 respectively over the baseline by incorporating the user profile information with off-the-shelf NLP features.

We also found several interesting observations. For example, the picture existence of the user can help to predict the quality of answer candidates as the first step. While the level-related information, i.e. level, points of user seems no influence on predicting the answer quality. Generally, the engagement-related information contributes more on answer ranking, followed by authority-related information.

As the AR information may have a bottleneck in the performance, we would like to incorporate some simple textual features to improve the robustness as well the performance of this method in the future. Moreover, we would investigate into the user model, especially the expertise model, to see how much it will help on answer ranking.

Acknowledgements

This research is supported by grants from National Natural Science Foundation of China (No.60903093) and Doctoral Fund of Ministry of Education of China (No.20090076120029).

7. REFERENCES

[1] J. Bian, Y. Liu, E. Agichtein, and H. Zha. Finding the right facts in the crowd: factoid question answering over social media. In *WWW*, pages 467–476, 2008.

[2] M. J. Blooma, A. Y.-K. Chua, and D. H.-L. Goh. A predictive framework for retrieving the best answer. In *SAC*, pages 1107–1111, 2008.

[3] M. Bouguessa, B. Dumoulin, and S. Wang. Identifying authoritative actors in question-answering forums: the case of yahoo! answers. In *KDD*, pages 866–874, 2008.

[4] J. Jeon, W. B. Croft, J. H. Lee, and S. Park. A framework to predict the quality of answers with non-textual features. In *SIGIR*, pages 228–235, 2006.

[5] P. Jurczyk and E. Agichtein. Discovering authorities in question answer communities by using link analysis. In *CIKM*, pages 919–922, 2007.

[6] J. Ko, L. Si, and E. Nyberg. A probabilistic framework for answer selection in question answering. In *HLT-NAACL*, pages 524–531, 2007.

[7] B. Li, I. King, and M. R. Lyu. Question routing in community question answering: putting category in its place. In *CIKM*, pages 2041–2044, 2011.

[8] M. Liu, Y. Liu, and Q. Yang. Predicting best answerers for new questions in community question answering. In *WAIM*, pages 127–138, 2010.

[9] Q. Liu and E. Agichtein. Modeling answerer behavior in collaborative question answering systems. In *ECIR*, pages 67–79, 2011.

[10] Q. Liu, E. Agichtein, G. Dror, E. Gabrilovich, Y. Maarek, D. Pelleg, and I. Szpektor. Predicting web searcher satisfaction with existing community-based answers. In *SIGIR*, pages 415–424, 2011.

[11] Y. Liu and E. Agichtein. You've got answers: Towards personalized models for predicting success in community question answering. In *ACL (Short Papers)*, pages 97–100, 2008.

[12] A. Moschitti and S. Quarteroni. Linguistic kernels for answer re-ranking in question answering systems. *Inf. Process. Manage.*, 47(6):825–842, 2011.

[13] L. Nie, B. D. Davison, and B. Wu. From whence does your authority come? utilizing community relevance in ranking. In *AAAI*, pages 1421–1426, 2007.

[14] S. Quarteroni. Personalized question answering. *TAL*, 51(1):97–123, 2010.

[15] C. Shah and J. Pomerantz. Evaluating and predicting answer quality in community qa. In *SIGIR*, pages 411–418, 2010.

[16] M. Surdeanu, M. Ciaramita, and H. Zaragoza. Learning to rank answers to non-factoid questions from web collections. *Computational Linguistics*, 37(2):351–383, 2011.

[17] X.-J. Wang, X. Tu, D. Feng, and L. Zhang. Ranking community answers by modeling question-answer relationships via analogical reasoning. In *SIGIR*, pages 179–186, 2009.

[18] J. Zhang, M. S. Ackerman, and L. A. Adamic. Expertise networks in online communities: structure and algorithms. In *WWW*, pages 221–230, 2007.