Analyzing Bias in CQA-based Expert Finding Test Sets

Reyyan Yeniterzi
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213
reyyan@cs.cmu.edu

Jamie Callan
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213
callan@cs.cmu.edu

ABSTRACT

Data retrieved from community question answering (CQA) sites, such as content and users' assessments of content, is commonly used for expertise estimation related tasks. One such task, in which the received votes are directly used as graded relevance assessment values, is ranking replies of a question. Even though these available assessments values are very practical for evaluation purposes, they may not always reflect the correct assessment value of the content, due to the possible temporal or presentation bias introduced by the CQA system during voting process. This paper analyzes a very commonly used CQA data collection in terms of these introduced biases and their effects on the experimental evaluation of approaches. A more bias free test set construction approach, which has correlated results with the manual assessments, is also proposed in this paper.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

Keywords

CQA sites; expert retrieval; test set construction

1. INTRODUCTION

Over the last few years, community question answering (CQA) sites became a common ground for knowledge sharing and acquisition among people. Additional to their every day use by the general community, data retrieved from these sites have been commonly used by the research community lately. The availability of user created content, social network structures and users' manual assessments of content through voting and best reply selection make these collections useful for many research problems. For example, research on expert retrieval [1] and related tasks [2, 3] often use data from CQA sites to rank answers based on their accuracy, relevance, or expertise. For assessing these expertise related tasks, previous work used questions as the test queries and the votes assigned to their replies as the ground truth assessments for ranking replies. No additional manual assessment was performed on replies or their authors. For test set construction, they either chose questions randomly or applied several restrictions such as choosing questions from a certain question category [2] or from a certain time frame [2], or choosing questions with responders who have replied to at least some number of questions [3].

However, test sets constructed with these approaches may contain several biases which may effect the reliability of the data and the experiments performed with it. Our research investigates two potential sources of bias in test set construction. The first one is a temporal bias caused by voting activities which are performed before receiving all answers. Another similar bias is the presentation bias which is caused by the initial ranking of the replies, when users only look at the top ranked replies, vote for one, ignore the rest of the replies and leave the page. These biases can lead to assessments of replies without actually viewing all the replies and therefore cause construction of incompetent test sets. This paper uses data retrieved from a popular CQA site and analyses these issues and investigates the effects of these introduced biases on the evaluation of approaches. Furthermore, this paper proposes a customized test set construction approach which is more bias-free and also more similar to the manual assessments both in terms of the way assessments are created and the results of the tested approaches.

2. DATASET

StackOverflow[1] is a CQA site focusing on technical topics such as operating systems, programming languages and environments. Users can post questions, reply to questions or leave comments to both questions and answers. A public data dump of StackOverflow site which contains all questions, answers and comments posted until June 2013, is used. Statistics related to this collection are provided in Table 1.

# Questions	5,130,696
# Answers	9,457,926
# Askers	936,864
# Responders	630,410
# Users	2,055,496

Table 1: Statistics of the StackOverflow collection.

[1]http://stackoverflow.com/

SIGIR'14, July 6–11, 2014, Gold Coast, Queensland, Australia.

http://dx.doi.org/10.1145/2600428.2609486.

Similar to other CQA sites, StackOverflow site also supports a voting mechanism in order to rate replies based on their accuracy and quality. Users can give 'up' or 'down' votes to replies, or askers can select a reply as the best. The number of votes given to each reply and the selected best replies are also available in the collection.

3. INTRODUCED BIASES

Our research investigates two potential sources of bias, temporal and presentation, in test set construction.

3.1 Temporal Bias

CQA systems display replies of questions as soon as they are posted, and they also let users vote for these replies as soon as they become online. Therefore, a reply may get votes even before other replies are posted. This paper initially analyzes whether users voting for answers before receiving all the answers to the question can be a problem or not. In order to test this, the percentage of replies, votes and (best reply) accepts received within certain time periods after the question is asked, are analyzed. The findings are presented in Table 2.

	Replies	Accepts	Votes
day 1	81.82%	50.84%	43.38%
days 2-7	5.54%	29.44%	10.33%
weeks 2-4	2.96%	9.12%	3.28%
weeks 5-52	5.79%	8.84%	15.65%

Table 2: Distribution of replying, voting and accepting best replies over time in certain time periods starting from the question posting time.

According to the percentages presented in Table 3, most of the replying (more than 80%) is performed within the day questions are asked but still 20% of the replies are still not provided within the first 24 hours. However, half of the accepts and around 40% of the votes are given in the first day, which indicates that part of the voting and accepting activities may have been performed before all the replies are received. A more detailed analysis showed that among the given accepts and votes, 5.79% of the accepts and 10.72% of the votes were given before the last reply was provided to the question. In such a system, replies posted earlier have higher probability to receive more votes or accepts, since they were available to users much earlier then the others. Therefore, data collections retrieved from CQA sites which let users vote for replies before receiving all the replies, can be biased towards replies that are posted earlier then the ones posted later.

3.2 Presentation Bias

Table 2 is also useful for understanding the possible effects of the presentation bias. As can be seen in table, only around half of the votes are given within the first week question is posted. The rest of the voting is performed much after, and even around 30% is received after a year is passed from the time question is asked. These votes are probably received from users who are directed to the question's page in CQA site by a search engine as a results of a search query. In such a web search like scenario, users may view the replies like in a way they view the web search engine results; starting from the top ranked reply they look through replies until they find what they are looking for, vote for it and leave the page without going over the remaining replies.

In such a user scenario, presentation order of the replies can affect the overall votes the replies can receive. Most CQA systems use their default reply ranking algorithms until replies receive votes from users. For instance, StackOverflow site ranks replies based on their posting time which means that when a user clicks on a question with replies that have not been voted yet, the system displays the earlier posted replies in top ranks and the later posted replies following them in order. Such a temporal ranking together with the only top ranked results viewing behavior of users, may lead the earliest submitted replies to receive more votes than others. In order to check this, the distribution of most voted replies are analyzed with respect to their posting times.

Reply Count	Percentage of Questions with Highest Voted Reply				
	First	Second	Third	Fourth	Fifth
2	65.9%	50.5%			
3	51.5%	41.7%	30.9%		
4	43.6%	34.6%	28.4%	19.9%	
5	37.9%	29.2%	24.9%	19.7%	13.7%

Table 3: Distribution of most voted replies (including ties) with respect to their posting times.

Table 3 reports the distribution of the highest voted reply for questions with 2 to 5 replies (including ties[2]). In this table, the column labeled as *first* presents the percentage of highest voted replies among all the first replies within that question category. According to the table, within questions with two replies, the first reply received the highest votes in around 65% of the time, while in an unbiased environment, it should have been around 50%, or distributed more to later replies, since they have the opportunity to improve on the previously posted replies. The decreasing percentage of highest voted replies from earlier replies to later replies in questions with 3-5 replies also indicates that voting is biased towards replies that were submitted earlier. This biased distribution of votes to top ranked replies strengthens the proposed hypothesis on presentation bias.

4. THE EFFECTS OF INTRODUCED BIASES

All these system features and user behaviors favor replies that are posted earlier. In order to check the effects of these biases on received votes, a manual assessment of best reply selection was performed on questions that have either their first or last reply received the highest votes from users. According to our hypothesis, if such voting related biases exist, then questions with first reply voted as highest should get lower agreement score between manual and voting based assessments, compared to questions with last reply voted as highest. This is mainly because, in questions with last reply received the highest vote, even though with the default ranking of the system, at least one user went through all the replies and voted for the latest submitted reply as best. Since a similar user behavior of viewing all replies is used in manual assessments, the agreement between manual and voting based assessments of replies should be higher in last

[2]Because we are including the tie cases, the sum of all percentages in a row is not equal to 100%

reply voted highest questions than the first reply voted highest ones.

4.1 Manual Assessments

In the manual assessment, questions that have either earliest or latest reply selected as best were displayed to the assessors after randomly sorting the replies. Assessors were asked to select the best reply based on its relevancy to the question, accuracy in the information provided and quality in expressing the answer. Assessors used these criteria and chose a reply as best without knowing either the original reply ID, responder ID, or reply posting time. In case the assessors believed that several replies are equally similar in terms of their relevance, accuracy and quality, they were given the freedom to choose multiple replies as best.

The assessments were performed by 7 volunteer assessors from our research group on randomly selected questions. A total of 300 questions with the highest voted replies equally distributed to first and last replies were used in the assessments. Assessors were given all the questions and asked to assess questions they feel comfortable with and as many of them possible in half an hour time. In order to increase the number of questions assessed in half an hour period, only questions with 2 to 4 replies were displayed in the assessment system.

A total of 98 questions were assessed by the assessors. Before performing any investigation on these assessments, a bias analysis was performed on the assessors and their selections in order to make sure that they are bias free. One assessor was identified as a possible biased case where he selected the first reply more than the sum of all the other ranks he selected in randomly sorted 2 to 4 reply questions. This assessor and his assessments were removed, due to the possible bias he may have introduced to the assessments.

Total of 86 questions remained in which 24 of them were assessed by multiple assessors. Among these 24 questions, the assessors agreed on the best reply on 17 of them while had different choices for the rest of the 7 questions. Within these 7 questions, the 3 of them were assessed by more than 2 assessors. Majority rule was applied to these 3 questions in order to select the more probable assessment value; however, this did not work on 4 questions which were assessed by only two assessors. In order to make the data more reliable and consistent, these 4 questions were removed from the test set which decreased the number of assessed questions to 82.

Even after removing these questions, there are still questions with multiple assessment values existed in the test data. These are the tie cases in which user could not select one reply as best and so selected several replies that have the same accuracy and quality. In order to analyze the results more clearly, two sets of results, one that excludes these ties (60 questions) and another one that includes them (82 questions) are provided.

The agreement ratio between these manual assessments and highest voted replies are summarized in Table 4. In an unbiased environment, such an assessment should return similar ratios for both first and last replies that received the highest votes. But as can be seen from the Table 4, the match ratio of the best selected replies are higher in last reply voted highest questions than the first reply voted highest questions for both with and without tie cases. These results strengthen the proposed hypothesis of existing bias between the earlier replies and votes assigned to them.

	SO Highest Voted Reply	# Manual Agree	# Manual Disagree	Agreement Ratio
Excl. Ties	First	16	20	0.44
	Last	14	10	0.58
Incl. Ties	First	25	24	0.51
	Last	20	13	0.61

Table 4: Agreement ratios between StackOverflow (SO) highest voted and manually assessed replies.

4.2 The Effects of These Introduced Biases on Expert Finding

In order to analyze the effects of these biases on expert finding related tasks, two widely used expertise estimation approaches for CQA environments were applied to test sets constructed with using only questions with either first or last replies received the highest vote. These two approaches are applied to rank the replies based on their authors' question specific expertise. In *Answer Count (AC)* [4] approach the replies are ranked by the number of topic relevant answers provided by the responder. In *Best Answer Count (BAC)* [2] approach, only the topic relevant answers that are selected as best are counted. In previous work, counting only the best selected replies is shown to be more effective than counting all replies of the responder [2].

In order to see the effects of tested approaches more clearly with respect to questions with different number of replies, a test set of 350 questions, which consists of randomly selected 50 questions with 2 to 7 replies, was constructed. Apart from this, no other selection criteria or restriction was applied in test set construction. During experiments, the question body was used as the query, and almost all (at most 5000) query relevant replies of the responder were retrieved for each question. The results of these experiments are summarized in Table 5.

Reply Count	First Reply		Last Reply	
	AC	BAC	AC	BAC
2-7	0.4467	**0.4700**	**0.3367**	0.3333
2	0.6800	**0.7000**	**0.5400**	0.5200
3	0.6200	**0.6400**	**0.4400**	0.3800
4	0.4600	**0.5000**	0.3200	**0.3600**
5	0.3600	**0.4400**	0.2800	**0.3000**
6	**0.4000**	0.3800	**0.2600**	0.2400
7	0.1600	0.1600	0.1800	**0.2000**

Table 5: Results for ranking replies based on their authors' question specific expertise.

Table 5 presents the experimental results of applying AC and BAC approaches to test sets in which only the first reply or the last reply get the highest votes. The first row presents the average best reply prediction accuracy results for all 350 questions while the rest of the rows present detailed results for each question category with different number of replies.

In Table 5, the *first reply* test set presents a similar behavior as reported in previous work [2], BAC performs better than the AC approach. This is expected because in randomly constructed test sets, the number of questions with earlier reply selected as best (or receives the highest votes) will be probabilistically higher since the distribution of most voted replies are higher in earlier submitted replies as shown

in Table 3. Similar results are also observed in questions with changing reply counts.

Unlike the *first reply* test set, the results from the *last reply* test set is different from the previously reported results. First of all, the results in *last reply* set are much lower than the results of *first reply* set, mainly because both algorithms applied are voting based algorithms which favor active users of the environment who are most likely to answer questions much quicker than other less active users and so have higher probability to be selected as best or voted as highest. Therefore, these voting based approaches may not be the best choice for identifying less active but more question specific expert responders.

Secondly, the relative ranking of the approaches are also different in *first reply* and *last reply* test sets. On average the AC approach slightly works better than the BAC and in terms of questions with different number of replies, there is not a consistent and clear winner. These results are different than *first reply* test set and previously reported results in literature [2] where there is a clear winner among the tested two approaches. This finding is especially important since such an inconsistency in results makes it hard to compare these two approaches and all other approaches accurately and reliably.

5. PROPOSED MORE BIAS-FREE TEST SET CONSTRUCTION APPROACH

In order to continue using these CQA data collections and the human assessments coming with them, in expert finding related tasks, this paper proposes a new test set construction approach which uses questions with later submitted replies selected as best or voted the highest. As mentioned before, in these type of questions, it is likely that at least one user assesses all the replies before selecting the best one. Such an approach is more bias-free and also more similar to the construction of controlled manual assessments.

In order to see whether the test set constructed with this approach is similar to the test set constructed with manual assessments, the *Answer Count* and *Best Answer Count* approaches were applied to the manually assessed best answer prediction test set (Section 4.1). Among the assessed questions in this test set, only 60 of them do not include any tie conditions and have clear winners, therefore, only these were used in the experiments. The results are presented in Table 6.

Reply Count	# Queries	AC	BAC
2-4	60	**0.4666**	0.4333
2	13	0.6154	0.6154
3	27	**0.4074**	0.3704
4	20	**0.4500**	0.4000

Table 6: Manual assessment results for ranking replies based on corresponding authors' expertise.

Table 6 contains the average results of the experiments over all 60 questions and as well as the more detailed results of questions with different reply counts. According to the table, overall the AC approach performs better than the BAC algorithm. Even though the number of questions in this experiment is far lower than the number of questions in the other previously tested test sets, the overall results and the performances of approaches look similar to *last reply* test set experimental results in Table 5. Especially, the results of questions with 3 replies, which has the highest number of questions (27), are very similar to the same category results of *last reply* test set (with 50 questions) in Table 5, which are 0.4400 instead of 0.4074 in AC and 0.3800 instead of 0.3704 in BAC approach.

These results strengthen the proposed idea of using questions with last reply selected as best in test set constructions, since their results are more similar to the results of manually assessed test sets. In StackOverflow, around 20% of the questions (among questions which has more than one reply) have their last reply received the highest votes and so can be used in the proposed test set construction approach. Of course one should be aware that this test set construction approach is proposed for CQA sites which use the posting time of replies in their default ranking algorithm until any user assessments (like votes) are received for replies. Similar but more customized approaches should be developed for specific CQA sites with different default reply ranking algorithms.

6. CONCLUSION

In this paper, a snapshot of StackOverflow site is used to analyze the widely used evaluation approach which uses user votes as the ground truth value. Two types of biases, temporal and presentation, are identified and their effects on the evaluation of the approaches are analyzed. It has been shown that these introduced biases affect the data in a way which is changing the relative performance of the applied approaches. Such a change in the relative ranking of approaches shows the significance of these biases, and their effects on the comparison of tested approaches.

In order to decrease the effects of these biases and create more accurate test collections, this paper proposes a test set construction approach that is customized for the tested CQA system's reply ranking algorithm. The similarity between the experimental results of the automatically constructed test set and the manually assessed test set indicates that the proposed approach can be used to create more bias-free test sets from these data collections without the need for doing manual assessments.

7. ACKNOWLEDGMENTS

This research was in part supported by National Science Foundation (NSF) grant IIS-1302206. Any opinions, findings, conclusions, and recommendations expressed in this paper are the authors' and do not necessarily reflect those of the sponsor.

8. REFERENCES

[1] K. Balog, Y. Fang, M. de Rijke, P. Serdyukov, and L. Si. Expertise retrieval. *Foundations and Trends in Information Retrieval*, 6(2–3), 2012.

[2] M. Bouguessa, B. Dumoulin, and S. Wang. Identifying authoritative actors in question-answering forums: The case of Yahoo! answers. In *Proceedings of KDD*, 2008.

[3] X. Liu, W. B. Croft, and M. Koll. Finding experts in community-based question-answering services. In *Proceedings of CIKM*, 2005.

[4] J. Zhang, M. S. Ackerman, and L. Adamic. Expertise networks in online communities: structure and algorithms. In *Proceedings of WWW*, 2007.