A Review paper on Exploiting User Feedback to Learn to Rank Answers in Q&A Forums: a Case Study with Stack Overflow

Siddhant Khandelwal
Department of Computer Science & Information Systems
Birla Institute of Technology & Science, Pilani (Pilani, India)
f20170127@pilani.bits-pilani.ac.in

Abstract—**Community question answering system is one of the fastest-growing users-generated-content systems. CQA enables users to ask/answer questions and search through the archived question-answers. CQA have made substantial progress in answering complex questions, such as mathematical or logical reasoning, subjective and open-ended questions. Being an open platform, CQA has little restrictions on who can post and who can answer a question. This review paper discusses the approach taken by the paper Dalip et al. [1] to rank answers in a popular CQA system, Stack Overflow. The paper suggests a Learning to Rank (L2R) approach for ranking answers in CQA systems. Particularly, the paper proposes a point-wise Random Forest supervised machine learning approach. The paper outperforms the then state-of-the-art approaches by 21% in NDCG, a metric used to evaluate rankings. This review paper also discusses the limitations in the authors' approach both in the erroneous underlying assumptions and lapses in the implementation. Future research directions pertaining to this specific problem are also discussed.**

Keywords—Community Question Answering Systems (CQA), Answer Quality, Answer Ranking, Learning to Rank

I. INTRODUCTION

In the present world, the importance of Community question-answering (CQA) systems, such as Yahoo! Answers or Stack Overflow is undeniable. These services are are used every day by billions of users to find an answer on varied questions, some may be complex, some purely subjective, or context-dependent questions. Therefore, for user satisfaction it is of utmost importance to present answers effectively. Therefore, CQA systems should utilize intelligence of its user base to provide with the most appropriate responses. As a result, CQA is a promising area of research in computer science. CQA systems result in creation of a large database of varying content quality. Many CQA systems adopt quality control mechanisms, where the users can help determine the relevance/quality of the answers and the quality of the question as well. This affects the reputation of the editor as well.

Answer quality is an important factor in CQA systems. Sometimes, the provided answers are not always appropriate, their quality can be diverse, and therefore better-quality answers can be get jumbled with inappropriate answers or even with spam and abusive content. The quality of an answer is usually determined by factors describing its content and relation to the question, such as accuracy, and comprehensiveness to the question, relevance, completeness and originality of the answer.

However, as scale of the system expands, it gets practically infeasible to rely on just the users to rate the answers. Moreover, these ratings are subject to bias and subjective factors like User expertise, relevance of background. In some cases, factors like tendency of abuse of some users might affect the rating of an answer too.

The problem of ranking answers in a CQA system thus is of utmost importance to the domain and calls for review of multiple state-of-the-art approaches. This review paper first presents research work prior to [1] and then describes the work of [1] in detail. The review paper then discusses the inherent limitations in the approach taken by the authors and the implementation limitations. Eventually, the paper focuses on alternate research directions to discuss the problem of answer ranking approaches.

II. PAST RESEARCH DIRECTIONS

Previous work has focused finding the best answers, rank these answers and assessing the quality of the question. Work is the first domain has been majorly a straight-forward classification task. A work of special attention by Shah et al. [2] which proposed nine different features in lieu of answer content and user information predicting the best answers in Yahoo Answers. The authors used a Logistic Regression based approach for this.

Work in the next domain, addressing the problem of ranking the answers and not just finding the best one is a little bit more involved. The work of special attention in this case is Suryanto et al. [3] achieved the best performance for this task. The authors proposed quality aware methods to rank answers. By using the proposed user expertise features, along with traditional relevance features, they were able to outperform all the previous work.

III. CONTRIBUTION OF THE PAPER

A. Problem Statement

CQA systems provide a way of loose edit control, which allows people to edit the content they post (or someone else posts) with mild restrictions. Due to this, the quality of the content raises much concerns. It becomes essential to control quality of this content. Some of these approaches are manual, however due to the scale of the CQA systems, it becomes extremely difficult as the scale increases. Moreover, the review of this content requires domain expertise which is highly subjective. This has a negative effect on various aspects of the CQA systems, mainly ranking and recommendation of content.

B. *Solution Approch*

To deal with the above problem, the paper in discussion proposes a Learning to Rank (L2R) approach to rank answers in a CQA system, in this case, popular CQA system called Stack Overflow. The paper proposes a pointwise L2RR method to assign a quality score for the answers.

From the many available approaches, the authors move ahead with Random Forest point-wise supervised machine learning approach to rank the answers. In a supervised learning approach, each query-document pair is represented by a set of features and annotated with a numerical score indicating how relevant is the given document to the query is. The features used are usually related to the similarity between the document and the query. A training set is therefore constructed to train the machine learning model which can then retrieve relevant documents pertaining to the query.

The authors choose to go ahead with a pointwise approach utilizing Random Forests for the machine learning model mainly since RF approaches are insensitive to parameter choices, resilient to overfitting and extremely parallel. In particular, as the decision tree regression algorithm, the authors use Multiple Additive Regression Trees, an implementation of the Gradient Tree Boosting method, provided by RankLib L2R Tool[4]. The algorithm proposed in the paper applies a decision tree regression algorithm to a subset of results multiple times and average the results. The process eliminates overfitting by using different subsets in each iteration from the same underlying distribution. The subsequent trees in each iteration are independent from others, thus making the RF trees inherently independent from each other.

The above described idea is directly used with CQA systems. The questions act as the queries and the answers as the document. Since most of the answers to a question in a CQA system are relevant, the authors focus on predicting a measure of answer quality rather than measuring the relevance of the answer.

The paper in total studies 189 features, out of which the paper innovates with 89 features. 98 of these total features are textual. The authors have grouped these features into *user, review, user graph, structure, length, style, readability and relevance.*

The user features directly infer the quality of the answer, by examining the user who has posted the answer. User graph features focus on the expertise level of the user who answers the question by examining the user's relationships with other users. Review features study the edit interactions of the posted answers as it is a well-known fact that an answer gets refined after peer-review and multiple edits and suggestions. Structure features describe the answers' organisation of content, presentation, usage of images, hyperlinks, codes snippets. The paper also involves features that focus on the length of the answers as an innovation. Style features capture the writing style of the answer, grammatical errors, sentence lengths, typos and language models used. Readability features like Smog-Grading, Gunning Fog Index are used to gauge the answers. Relevance features like BM25 ranking function, common words, sentences among others are used as suggested in previous studies.

C. *Dataset*

The authors' dataset consists of a sample of Stack Overflow, a Q&A Forum for programmers. Stack Overflow is the largest collection of Stack Exchange and is freely available to download. Stack Overflow has fewer spam and distractions when compared to general Q&A Forums such as Yahoo Answers.

The authors extracted only questions that have, at least, 4 answers. Questions with less than 4 answers can easily be assessed by the users. Authors also removed questions with no rated answers. The dataset has 9,721 questions with 53,263 answers. For the creation of the user graph, authors have considered all the users of Stack Overflow.

D. *Learning function*

The paper uses the difference between upvotes and downvotes received by the answer referred to as the answer rating, given by:

$$r_a = r'_a + r'_{min}$$

where $r_a^{`} = u_a - d_a$ is the difference between the number of upvotes $u_a$ and downvotes $d_a$ received by answer $a$, and $r_{min}^{`}$ is the minimum difference between upvotes and downvotes observed in the collection.

E. *Baselines*

The baselines adopted in the paper are slightly modified versions of the methods proposed by Suryanto et al [3], Shah et al [2], and Burel et al [5]. The authors refer to their methos as RF, to the method proposed by Suryanto et al as EXQD, to the method proposed by Shah et al as SHAH, and to the method proposed by Burel et al as AdTree. The RF approach uses all the features previously described. In order to test the effectiveness of the new features proposed, the authors implemented a version of RF where Q&A pairs are represented only with features used by the baselines (the ones that are not new). The authors refer to this method as RF-BaseFeatures. The authors modified the SHAH baseline to not just predict the best answer, but also to rate the answers based on a logistic regression variant of the learning function above.

F. *Tools used for features*

To compute some style and readability features the authors used the Style and Diction software[8]. Vocabular features used the English lexical database WordNet[9]. The count of typos in the answers was calculated using the words present in a list of common misspellings, available from Wikipedia[10].

G. *Evaluation Metric*

The author use the metric *Normalized Discounted Cumulative Gain at top k (NDCG @k).* It helps to measure how close the predicted answer is to the ground truth ranking.

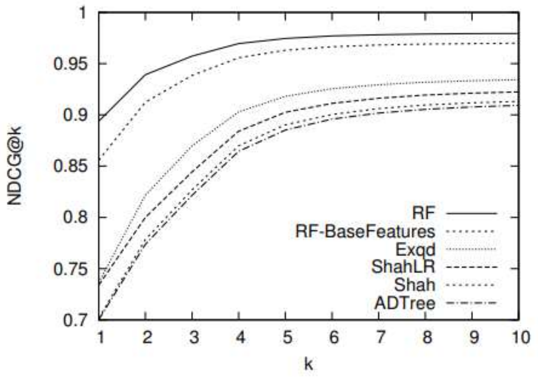
$$NDCG@k = \frac{1}{N}\sum_{i=1}^{k}(\frac{2^{r_i}}{\log_2(i+1)})$$

Here, r_i is the true rating assessment for the answer at position i in the ranking, and N is a normalization factor. The factor N is equal to the discounted cumulative gain of the ideal ranking, i.e. the ranking where, given a pair of answers (a_i, a_j), a_i is better ranked than a_j if r_i` is greater than $r_j$`.

For all the comparisons reported in the paper, the authors have used the signed-rank test of Wilcoxon to determine if the differences in the effectiveness were statistically significant. Apart from this, the authors have used Infogain metric to evaluate the impact of the chosen features. Infogain is a statistical measure of how much a given feature contributes to discriminate the class to which any given article belongs.

H. Results

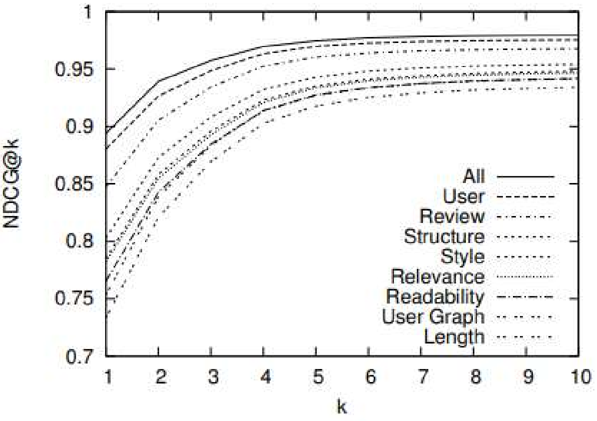
The paper tests the effectiveness of the new features propose by implementing a version of RF with only the already used features in past work and one with the old as well as the new features. The RF with the new features outperformed all the previous implementations and the one with out the new features too, over all values of *k*. All the differences pointed out between RF and the other methods were statistically significant, at all *NDCG@k* points, according to the Wilcoxon test.



The RF approach achieves gains ranging from 6% (*NDCG@10*) to 21% (*NDCG@1*).

To analyse the performance of the group of features, the features were divided into 8 groups and subsequently, two series of experiments were conducted. One with a group alone and the other set of experiments with only one category excluded in each iteration. This way, the authors were able to determine how each group fared.

When the groups were taken in isolation none of the group performed significantly better than others. However, User group features turned out to be the most relevant. This is also reflected when this group was excluded in an iteration. The next best set being the features of the review group.



Relevance features have no impact probably due to the fact that most of the answers in Stack Overflow are relevant to their questions. Readability features are probably not suitable for this scenario, characterized by short text and many code snippets.

Length features are probably redundant since many other features (eg, number of prepositions) are correlated to length. User Graph features presented a bad performance because they failed to capture user expertise in the dataset.

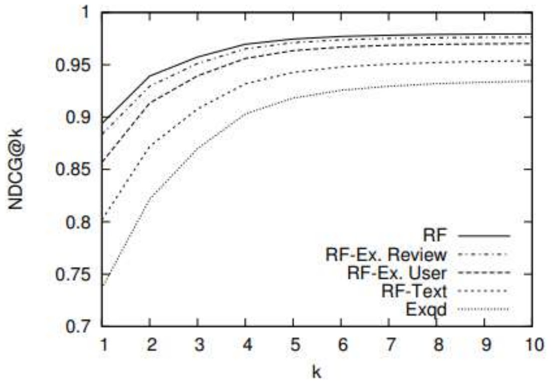
Using the infogain metric, the conclusions confirmed the good performance of user and review features.

Group	# of features at top...					
	1-10	11-20	21-30	31-40	41-50	>50
User	7	8	2	3	3	36
Review	3	2	3	0	1	9
Structure	0	0	1	0	1	23
Length	0	0	2	1	0	0
Style	0	0	2	1	4	23
Relevance	0	0	0	5	1	34
Readability	0	0	0	0	0	7
User Graph	0	0	0	0	0	4

I. Rating new answers and new users

It is important to note that not all the feature groups are always available to rank the answers. New answers may not have any relevant Review features. Similarly, new users may not have any relevant User features. Therefore, the authors conducted experiments to cover these cases.

Even without User or Review features, the authors obtained very high *NDCG@k* values, close to those obtained with all features. Using only text features, results visibly decrease. However, they are still well above the EXQD baseline, thus showing the strength of this method, even when using much less information.



IV. LIMITATIONS

- User Graph features performed badly on the dataset. Diversifying the study to other CQA systems where the relationship between the users is more influential is necessary to decide whether to disregard the feature or not.

- The method to obtain ground-truth, i.e. the difference between the number of upvotes and downvotes is prone to bias and needs expert review.

- The paper does not discuss about questions that do not have a clear-cut top-rank answer. Some questions can have multiple correct answers with none of them being truly complete. These answers can be representing different perspectives to the same question. This is applicable even with programming questions as is the case with this paper. Several concepts in programming have this notion of multiple perspectives with respect to different implementations of the concept. There is not a clear top-rank answers in such situations and these types of cases need to be handled differently.

- Pointwise approaches of ranking deal with the problem as classification or regression on single QA pairs. They are thus unable to consider the relative orders of two answers. This results in sub-optimal performance.

V. FUTURE RESEARCH DIRECTIONS

- Another direction of research could be to better involve user information features in order to influence ranking decisions.

- Topic statistics of the question can be better utilized as a feature for the model. Different topics might have different ways of answering questions, different language models, or presentation structure of the answer. Experts of these topics can be utilized to identify specificities in these topics and proper ranking models can be constructed for the same.

- Novelty detection is a statistical method used to determine new or unknown data and determining if these new data are within the norm (inlier) or outside of it (outlier) [6].

- Novelty-based ranking approaches that take into consideration the wider scope of answers and different perspectives highlighted therein can be another direction of research. Adopting to pairwise or listwise approaches will result in better results with the right set of features for the machine learning model as suggested in Omari et al [7].

ACKNOWLEDGMENT

This review paper is written in order to complete the course requirements for Information Retrieval course under the guidance of Prof. Abhishek of BITS Pilani, Pilani Campus.

REFERENCES

[1] D. H. Dalip, M. A. Gonc¸alves, M. Cristo, and P. Calado, “Exploiting user feedback to learn to rank answers in QA forums: A case study with stack overflow,” in Proc. 36th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2013, pp. 543–552.

[2] C. Shah and J. Pomerantz, “Evaluating and predicting answer quality in community QA,” in Proc. 33rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2010, pp. 411–418.

[3] M. A. Suryanto and R. H. L. Chiang. Quality-Aware Collaborative Question Answering : Methods and Evaluation. In WSDM ’09, pages 142–151, 2009.

[4] http://people.cs.umass.edu/~vdang/ranklib. html

[5] Automatic identification of best answers in online enquiry communities. In G. Burel, Y. He, and H. Alani, editors, 9th Extended Semantic Web Conference, volume 7295 of Lecture Notes in Computer Science, Crete, 2012. Springer Berlin Heidelberg.

[6] https://www.techopedia.com/definition/30345/novelty-detection

[7] Omari, Adi, et al. "Novelty based ranking of human answers for community questions." *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. 2016.

[8] http://www.gnu.org/software/diction/

[9] http://wordnet.princeton.edu

[10] http://en.wikipedia.org/wiki/Wikipedia:Lists_of_common_misspellings