Survey Paper on a Pairwise Learning to Rank Model for Answer Selection in Community Question Answer (CQA) System

Rahul Patil
Department of Computer Engineering,
Pimpri Chinchwad College of Engineering, India

Sunny Shah
Department of Computer Engineering,
Pimpri Chinchwad College of Engineering, India

Tejas Bavaskar
Department of Computer Engineering,
Pimpri Chinchwad College of Engineering, India

Sourabh Ukhale
Department of Computer Engineering,
Pimpri Chinchwad College of Engineering, India

Akash Kalyankar
Department of Computer Engineering,
Pimpri Chinchwad College of Engineering, India

ABSTRACT

To find the similar questions is very difficult in Question Answering (QA) System. Because each question in the returned candidate pool consists of multiple answers, and hence users get trouble to browse a lot before finding the correct one. To overcome this problem, we construct a novel approach a novel Pair wise learning to rank model i.e PLANE which can quantitatively rank answer candidates from the relevant question pool. Specifically, it comprises two components i.e. one offline learning component and one online search component. In the offline learning component, we first consequently set up the positive, neutral, and negative training samples in the forms of preference pairs guided by our data-driven observations. We at that point display a novel model to together consolidate these three sorts of preparing tests and the closed-form solution of this model is determined. In the online search component, we initially gather a pool of answer candidates for the given question by means of discovering its comparable or similar questions. We at that point sort the appropriate answer candidates by utilizing the offline trained model to judge the preference orders. We also design recommendation system, in which best solution is recommended. The system also provides facilities like bookmarking as well as sends best answer on email. Our model is robust as well as achieves better performance than several state-of-the-art answer selection baselines.

Keywords

Answer Selection, Community-based Question Answering, Naive Byes, pairwise learning, Recommendation.

1. INTRODUCTION

1.1. Background

Community Response System Questions (CQA), one User Generated Content (UGC) is fastest growing, Portals has risen as a huge market, so to speak, to meet the needs of complex information. CQA It allows users to ask/answer questions and search through historical question-answer (QA) Couple Compared to the traditional QAs in fact, like "Who is the President of Singapore in 2016". You can simply answer by removing the names bodies or paragraphs of documents, CQA Substantial progress in response to Questions, such as reasoning, open-ended and advance seeking questions. CQA is therefore completely open and little restrictions, where applicable, about who can publish and who can answer a question. Despite the success of CQA and active user Demand for famine participation exists extensively in CQA sites, which refers to the following two types Phenomena: 1. Firstly, those looking for information usually have to wait long before getting answers to their questions. 2. Secondly, a large percentage of questions do not any response even within a relatively long period. Question starvation is probably caused by several Reasons: 1. The questions are badly formulated, Ambiguous or less interesting to all. 2. CQA systems are hard to route the newly posted questions to the appropriate answerers 3. Potential interviewees have experience, but they are not available or overwhelmed by the pure volume of incoming questions. This case often occurs in the vertical CQA forums, so only experts can answer these questions.

1.2. Motivations

The main motivation is to overcome the problem of to find the similar questions, because each question in the returned candidate pool consists of multiple answers, and hence users get trouble to browse a lot before finding the correct one. So, we motivate to construct a novel approach a novel Pairwise Learning to rANk model i.e. PLANE which can quantitatively rank answer candidates from the relevant question pool.

1.3. Goal

The goal of our system is to achieve better performance as well as robustness through a novel Pairwise Learning to rANk model. i.e. PLANE.

1.4. Objective and Scope

- To achieve better performance.
- To provide security.
- To achieve robustness.
- To make system user-friendly.

2. RELATED WORK OR LITERATURE SURVEY

[1] W. Wei, Z. Ming, L. Nie, G. Li, J. Li, F. Zhu, T. Shang, and C. Luo (2016) have represented Exploring heterogeneous features for query-focused summarization of categorized community answers [1]. The author proposes a three-level scheme, which aims to generate a query-focused summary-style answer in terms of two factors, i.e., novelty and redundancy. Specifically, we first retrieve a set of Qas to the given query, and then develop a smoothed NaiveBayes model to identify the topics of answers, by exploiting their associated category information.

[2] X. Li, Y. Ye and M. K. Ng (2016) have represented Multivcrank with applications to image retrieval [2]. The author proposes and develops a multi-visual concept ranking (MultiVCRank) scheme for image retrieval. The key idea is that an image can be represented by several visual concepts, and a hypergraph is built based on visual concepts as hyperedges, where each edge contains images as vertices to share a specific visual concept. In the constructed hypergraph, the weight between two vertices in a hyperedge is incorporated, and it can be measured by their affinity in the corresponding visual concept. A ranking scheme is designed to compute the association scores of images and the relevance scores of visual concepts by employing input query vectors to handle image retrieval.

[3] W. Wei, G. Cong, C. Miao, F. Zhu, and G. Li (2016) has represented learning to find topic experts in Twitter via different relations [3]. The author develops a probabilistic method to jointly exploit three types of relations (i.e., follower relation, user-list relation, and list-list relation) for finding experts. Specifically, propose a Semi-Supervised Graph-based Ranking approach (SSGR) to offline calculate the global authority of users. In SSGR, employ a normalized Laplacian regularization term to jointly explore the three relations, which is subject to the supervised information derived from Twitter crowds. Then online compute the local relevance between users and the given query. By leveraging the global authority and local relevance of users, we rank all of the users and find top-N users with highest ranking scores.

[4] W. Wei, B. Gao, T. Liu, T. Wang, G. Li, and H. Li (2016) has designed A ranking approach on a large-scale graph with multidimensional heterogeneous information [4]. The author addresses the large-scale graph-based ranking problem and focuses on how to effectively exploit rich heterogeneous information of the graph to improve the ranking performance. Specifically, propose an innovative and effective semi-supervised PageRank (SSP) approach to parameterize the derived information within a unified semisupervised learning framework (SSLF-GR), then simultaneously optimize the parameters and the ranking scores of graph nodes.

[5] X. Wei, H. Huang, C. Lin, X. Xin, X. Mao, and S. Wang(2015) have represented Reranking voting-based answers by discarding user behavior biases [5]. In generating a vote, a user's attention is influenced by the answer position and appearance, in addition to real answer quality. Previously, these biases are ignored. As a result, the top answers obtained from this mechanism are not reliable, if the number of votes for the active question is not sufficient. The author solves this problem by analyzing two kinds of biases; position bias and appearance bias. To identify the existence of these biases and propose a joint click model for dealing with both of them.

[6] Q. H. Tran, V. Duc, Tran, T. T. Vu, M. L. Nguyen, and S. B. Pham (2015) Jaist: Combining multiple features for answer selection in community question answering [6].The author designed Answer Selection in Community Question Answering. In this task, the systems are required to identify the good or potentially good answers from the answer thread in Community Question Answering collections. This system combines 16 features belong to 5 groups to predict answer quality. This final model achieves the best result in subtask A for English, both in accuracy and F1-score.

[7] Savenkov (2015) has represented Ranking answers and web passages for non-factoid question answering: Emory University at TREC live QA [7]. The author represents how to automatically answer questions posted to Yahoo! Answers community question answering website in real-time. This system combines candidates extracted from answers to similar questions previously posted to Yahoo! Answers and web passages from documents retrieved using a web search. The candidates are ranked by a trained linear model and the top candidate is returned as the answer. The ranking model has trained on question and answer (QnA) pairs from Yahoo! Answers archive using pairwise ranking criterion. Candidates are represented with a set of features, which includes statistics about candidate text, question term matches and retrieval scores, associations between question and candidate text terms and the score returned by a Long Short-Term Memory (LSTM) neural network model.

[8] L. Nie, Y. Zhao, X. Wang, J. Shen, and T. Chua (2014) has represented Learning to recommend descriptive tags for questions in social forums [8]. Around 40% of the questions in the emerging social-oriented question answering forums have at most one manually labeled tag, which is caused by incomprehensive question understanding or informal tagging behaviors. The incompleteness of question tags severely hinders all the tag-based manipulations, such as feeds for topic-followers, ontological knowledge organization, and other basic statistics. The author presents a novel scheme that is able to comprehensively learn descriptive tags for each question.

[9] Z. Ji and B. Wang(2013) has represented Learning to rank for question routing in community question answering [9]. The author proposes a general framework based on the learning to rank concepts for QR. Training sets consist of triples (q, asker, answerers) are first collected. Then, by introducing the intrinsic relationships between the asker and the answerers in each CQA session to capture the intrinsic labels/orders of the users about their expertise degree of the question q, two different methods, including the SVM-based and Ranking SVM-based methods, are presented to learn the models with different example creation processes from the training set. Finally, the potential answerers are ranked using the trained models.

[10] H. Dalip, M. A. Gonc¸alves, M. Cristo, and P. Calado (2013) has represented Exploiting user feedback to learn to rank answers in qa forums: A case study with stack overflow [10].The author proposes a learning to rank (L2R) approach for ranking answers in Q&A forums. In particular, we adopt an approach based on Random Forests and represent query and answer pairs using eight different groups of features. Some of these features are used in the Q&A domain for the first time. The L2R method was trained to learn the answer rating, based on the feedback users give answers in Q&A forums.

[11] T. C. Zhou, M. R. Lyu, and I. King (2012) A classification based approach to question routing in community question answering [11]. The author designed a new approach to Question Routing, which aims at routing questions to participants who are likely to provide answers. The author considers the problem of question routing as a classification task and develops a variety of local and global features which capture different aspects of questions, users, and their relations.

[12] A Agarwal, H. Raghavan, K. Subbian, P. Melville, R. D. Lawrence, D. C. Gondek, and J. Fan(2012) have represented Learning to rank for robust question answering [12]. The main aims to solve the problem of improving the ranking of answer candidates for factoid based questions in a state-of-the-art Question Answering system. The author first provides an extensive comparison of 5 ranking algorithms on two datasets – from the Jeopardy quiz show and a medical domain. Then show the effectiveness of a cascading approach, where the ranking produced by one ranker is used as input to the next stage. The cascading approach shows sizeable gains on both datasets. Then finally evaluate several rank aggregation techniques to combine these algorithms, and find that Supervised Kemeny aggregation is a robust technique that always beats the baseline ranking approach used by Watson for the Jeopardy competition.

[13] Hieber and S. Riezler(2011) has developed Improved answer ranking in social question-answering portals [13]. Community QA portals provide an important resource for non-factoid question-answering. The inherent noisiness of user-generated data makes the identification of high-quality content challenging but all the more important. The author presents an approach to answer ranking and show the usefulness of features that explicitly model answer quality. Furthermore, introducing the idea of leveraging snippets of web search results for query expansion in answer ranking. Then present an evaluation setup that avoids spurious results reported in earlier work.

[14] B. Li and I. King (2010) has represented Routing questions to appropriate answerers in community question answering services [14]. Community Question Answering (CQA) service provides a platform for increasing number of users to ask and answer for their own needs but unanswered questions still exist within a fixed period. To address this, the main aims to route questions to the right answerers who have a top rank in accordance with their previous answering performance. In order to rank the answerers, the author proposes a framework called Question Routing (QR) which consists of four phases: (1) performance profiling, (2) expertise estimation, (3) availability estimation, and (4) answerer ranking.

[15] K. Wang, Z. Ming, and T.-S. Chua (2009) has represented A syntactic tree matching approach to finding similar questions in community-based QA services [15]. The author proposes a new retrieval framework based on syntactic tree structure to tackle the similar question matching problem. Then build a ground-truth set from Yahoo! Answers and experimental results show that our method outperforms traditional bag-of word or tree kernel-based methods by 8.3% in mean average precision. It further achieves up to 50% improvement by incorporating semantic features as well as matching of potential answers

3. EXISTING SYSTEM AND DISADVANTAGES

In the existing system, the question starvation widely exists in cQA system which refers to the following two types Phenomena: 1. Firstly, those looking for information usually have to wait long before getting answers to their questions. 2. Secondly, a large percentage of questions do not any response even within a relatively long period. Question starvation is probably caused by several Reasons: 1. The questions are badly formulated, Ambiguous or less interesting to all. 2. cQA systems are hard to route the newly posted questions to the appropriate answerers 3. Potential interviewees have experience, but they are not available or overwhelmed by the pure volume of incoming questions. This case often occurs in the vertical cQA forums, so only experts can answer these questions.

Disadvantages

- Time-consuming process.
- Difficulty to find a similar question.
- The ranking is not proper.

4. PROPOSED SYSTEM AND ADVANTAGES

The proposed system, construct a novel Pairwise Learning to rANk model i.e PLANE which can quantitatively rank answer candidates from the relevant question pool. Specifically, it comprises two components i.e one offline learning component and one online search component. In the offline learning component, we first consequently set up the positive, neutral, and negative training samples in the forms of preference pairs guided by our data-driven observations. We at that point display a novel model to together consolidate these three sorts of preparing tests and the closed-form solution of this model is determined. In the online search component, we initially gather a pool of answer candidates for the given question by means of discovering its comparable or similar questions. We at that point sort the appropriate answer candidates by utilizing the offline trained model to judge the preference orders.

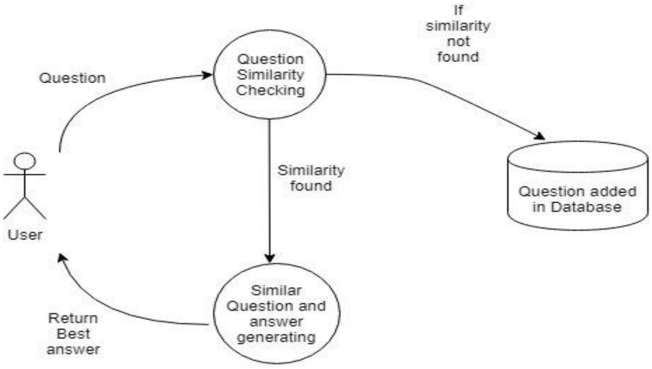
If similarity not found
Question
Question Similarity Checking
User
Similarity found
Question added in Database
Return Best answer
Similar Question and answer generating

Fig 1: Proposed System

Advantages

- Not time-consuming process.
- Achieve better performance.
- Provide security.
- Achieve robustness.
- Make system user-friendly

5. CONCLUSION

We present a novel scheme for answer selection in cQA system. It consists of one offline learning and the online search component. In component offline learning, instead of time-consuming and labor-intensive annotation, automatically builds positive, neutral and Negative training samples in the forms of guided by our observations on the database. We then propose robust pairwise learning to rank model to incorporate these three types of training samples. In the online search component, A particular question is, first of all, gathering a group of answers find candidates through their similar questions. We then use the offline model to classify candidate answers through pairwise comparison. We have conducted extensive experiments to justify the effectiveness of our model in a cQA general data set and a series of vertical cQA data.

We can conclude the Following points: 1. Our model can achieve better performance than several state-of-the-art answer selection baselines. 2. Our model is not sensitive to its parameters. 3. Our model is robust to noise caused by the expansion of applications. 4. Learn how to classify models in pairs including our proposed plan are very sensitive to the error training samples. We also design recommendation system, in which best solution is recommended. The system also provides facilities like bookmarking as well as sends best answer on email.

6. ACKNOWLEDGMENTS

Sunny Shah, Tejas Bavaskar, Sourabh Ukhale, Akash Kalyankar are the respective undergraduate students with the Department of Computer Engineering, Pimpri Chinchwad College of Engineering, India (Email: sunnyshah0588, tejkar777, Sourabh.ukhale, akalyankar12) (@gmail.com).This paper is a part of the project for their final year, submitted towards the Savitribai Phule Pune University (formerly University of Pune).

7. REFERENCES

[1] W. Wei, Z. Ming, L. Nie, G. Li, J. Li, F. Zhu, T. Shang, and C. Luo, “Exploring heterogeneous features for query-focused summarization of categorized community answers,” Inf. Sci., vol. 330, pp. 403–423, 2016.

[2] X. Li, Y. Ye, and M. K. Ng, “Multivcrank with applications to image retrieval,” TIP, vol. 25, no. 3, pp. 1396–1409, 2016.

[3] W. Wei, G. Cong, C. Miao, F. Zhu, and G. Li, “Learning to find topic experts on twitter via different relations,” TKDE, vol. 28, no. 7, pp. 1764–1778, 2016

[4] W. Wei, B. Gao, T. Liu, T. Wang, G. Li, and H. Li, “A ranking approach on a large-scale graph with multidimensional heterogeneous information,” TOC, vol. 46, no. 4, pp. 930–944, 2016.

[5] X. Wei, H. Huang, C. Lin, X. Xin, X. Mao, and S. Wang, “Reranking voting-based answers by discarding user behavior biases,” in Proceedings of IJCAI’15, 2015, pp. 2380–2386.

[6] Q. H. Tran, V. Duc, Tran, T. T. Vu, M. L. Nguyen, and S. B. Pham, “Jaist: Combining multiple features for answer selection in community question answering,” in Proceedings of SemEval’15. ACL, 2015, pp. 215C–219.

[7] Savenkov, “Ranking answers and web passages for non-factoid question answering: Emory university at TREC liveqa,” in Proceedings of TREC’15, 2015.

[8] L. Nie, Y. Zhao, X. Wang, J. Shen, and T. Chua, “Learning to recommend descriptive tags for questions in social forums,” TOIS, vol. 32, no. 1, p. 5, 2014.

[9] Z. Ji and B. Wang, “Learning to rank for question routing in community question answering,” in Proceedings of CIKM’13. ACM, 2013, pp. 2363–2368

[10] D. H. Dalip, M. A. Gonc¸alves, M. Cristo, and P. Calado, “Exploiting user feedback to learn to rank answers in qaforums: A case study with stack overflow,” in Proceedings of SIGIR’13. ACM, 2013, pp. 543–552.

[11] T. C. Zhou, M. R. Lyu, and I. King, “A classificationbased approach to question routing in community question answering,” in Proceedings of WWW’12. ACM, 2012, pp. 783–790.

[12] A. Agarwal, H. Raghavan, K. Subbian, P. Melville, R. D. Lawrence, D. C. Gondek, and J. Fan, “Learning to rank for robust question answering,” in Proceedings of CIKM ’12. ACM, 2012, pp. 833–842.

[13] Hieber and S. Riezler, “Improved answer ranking in social question-answering portals,” in Proceedings of SMUC’11.ACM, 2011, pp. 19–26.

[14] B. Li and I. King, “Routing questions to appropriate answerers in community question answering services,” in Proceedings of CIKM’10. ACM, 2010, pp. 1585–1588.

[15] K. Wang, Z. Ming, and T.-S. Chua, “A syntactic tree matching approach to finding similar questions in community-based qa services,” in Proceedings of SIGIR’09. ACM, 2009, pp. 187–194.

[16] M. A. M. L. Wei Ding, He Jiang, Modern Advances in Intelligent Systems and Tools, ser. SCI. Springer, 2012, vol. 431.

[17] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne, “Finding high-quality content in social media,” in Proceedings of WSDM’08. ACM, 2008, pp. 183–194.

[18] J. Jeon, W. B. Croft, J. H. Lee, and S. Park, “A framework to predict the quality of answers with non-textual features,” in Proceedings of SIGIR’06. ACM, 2006, pp. 228–235.

[19] Z. Ji and B. Wang, “Learning to rank for question routing in community question answering,” in Proceedings of CIKM’13. ACM, 2013, pp. 2363–2368.

[20] T. C. Zhou, M. R. Lyu, and I. King, “A classification based approach to question routing in community question answering,” in Proceedings of WWW’12. ACM, 2012, pp. 783– 790.

[21] L. Yang, M. Qiu, S. Gottipati, F. Zhu, J. Jiang, H. Sun, and Z. Chen, “Cqarank: Jointly model topics and expertise in community question answering,” in Proceedings of CIKM’13. ACM, 2013, pp. 99–108.

[22] B. Li and I. King, “Routing questions to appropriate answerers in community question answering services,” in Proceedings of CIKM’10. ACM, 2010, pp. 1585–1588.