Postprint: Answer Ranking in Social Q&A Communities Under Multi-dimensional Features

Authors: Yiming, Zhang Tingting, Li Zi

Date: 2023-04-01T16:16:00+00:00

Abstract

[Purpose/Significance] This study investigates the impact of multi-dimensional features on answer ranking in social Q&A communities, aiming to improve community service quality and optimize user experience. [Method/Process] We construct a feature system for answer ranking in social Q&A communities from three dimensions: answer features, answerer features, and voter features. The applicability of 11 learning-to-rank algorithms based on deep learning, tree models, neural networks, support vector machines, etc., is compared on Q&A community datasets, and a Random Forest classification algorithm is trained to obtain the importance of each feature. [Results/Conclusion] Experimental results demonstrate that deep learning-based learning-to-rank algorithms outperform other ranking algorithms on both NDCG@k and MRR metrics. Voter influence features are the most important, followed by answer content features, and finally answerer expertise features. Further optimization of answer ranking can be considered from two dimensions: increasing the diversity of answer ranking methods and enhancing the comprehensiveness of answer ranking algorithms.

Full Text

Preamble

Research on Answer Ranking in Social Q&A Communities Based on Multidimensional Features

Yi Ming, Zhang Tingting, Li Ziqi

School of Information Management, Central China Normal University, Wuhan 430079

Abstract: [Purpose/Significance] This study investigates the impact of multidimensional features on answer ranking in social Q&A communities to improve service quality and optimize user experience. [Method/Process] We con-

structed a feature system for answer ranking from three dimensions: answer features, respondent features, and voter features. We compared the applicability of 11 learning-to-rank algorithms (including deep learning, tree-based, neural network, and support vector machine methods) on Q&A community datasets, and trained a random forest classification algorithm to determine the importance of each feature. [Result/Conclusion] Experimental results show that deep learning-based ranking algorithms outperform other methods on NDCG@k and MRR metrics. Voter influence features are most important, followed by answer content features, and finally respondent expertise features. We propose optimizing answer ranking by increasing ranking method diversity and improving algorithm comprehensiveness.

Keywords: Social Q&A Community; Answer Quality; Learning-to-Rank Algorithm; Deep Learning Algorithm

In the Web 2.0 era, social Q&A communities have become important online platforms for knowledge acquisition and interactive communication. After more than a decade of development, these communities have matured and profoundly influenced users' knowledge acquisition and social behaviors [1]. However, alongside the convenience of interaction, issues such as "knowledge overload" have emerged. Research on answer ranking in social Q&A communities is crucial for providing high-quality answers and improving service quality, and has become a key focus in both practice and research.

Current research on answer ranking primarily examines three dimensions: answer structural features, answer text features, and respondent features, but the selected feature dimensions are not comprehensive. Most studies concentrate on algorithm optimization, resulting in numerous ranking algorithms but lacking evaluation of different algorithm categories' applicability to social Q&A community datasets. Moreover, given the large user base and diverse information needs, existing ranking methods struggle to satisfy all users. For instance, some users prefer expert answers while others may prefer content-rich or strongly subjective/objective answers. This paper addresses these limitations by constructing a multidimensional answer ranking feature system and integrating it into ranking algorithms to improve service quality and user experience.

2 Research Status of Answer Ranking in Social Q&A Communities

2.1 Feature Indicators for Answer Ranking

Existing research primarily uses answer and respondent features for ranking, with few studies examining voter characteristics.

2.1.1 Answer Features Answer features include external and internal characteristics. External features are directly measurable: (1) answer length [2-4]; (2) number of links and images, where more such content indicates richer answers

[2-5]; (3) upvotes, comments, downvotes, and views, where higher participation suggests more popular and likely higher-quality answers [5]; (4) counts of nouns, verbs, and interrogative words, as well-structured answers with appropriate syntactic features are more likely to be good answers [2-3,6]; (5) number of sentences [7]; and (6) other features like answer-to-question ratio [2] and identical word sequences [6,8].

Internal features are embedded in text and not directly observable: (1) question-answer similarity—higher thematic similarity indicates higher quality [2,4-5,7]; (2) answer similarity—calculating similarity among answers to the same question can filter irrelevant answers [9-10]; and (3) sentiment polarity—answers with positive sentiment 倾向 are more likely to be good answers [1-2].

2.1.2 Respondent Features Respondent features measure expertise, variously termed user authority, professionalism, or expertise [3-5,8,11-14]. We uniformly refer to this as user expertise. Researchers measure expertise through: disclosed personal specialties and follow domains [12]; more answers and fewer questions in a topic indicating higher expertise [5]; normalized best answer counts and accuracy rates in a domain [4]; and question type impact through classification models [15]. Answer quality also relates to response time—high-quality answers take longer, mobile responses are faster but lower quality, and anonymous responses are faster but rarely high-quality [16].

2.1.3 Voter Features Geerthik et al. [17] constructed an effective ranking model from voter characteristics, including: respondent follower count, upvotes from followers/non-followers, expert upvotes/downvotes, upvotes from other respondents, and non-follower downvotes. Cui et al. [18] used feature-based Borda Count to fuse multi-evaluation criteria.

2.2 Answer Ranking Methods

2.2.1 Using Existing Ranking Models Some researchers combine theoretical approaches with existing models. Surdeanu et al. [6] incorporated question-answer similarity, term density/frequency, and network features into perceptron and SVMRank models. Yuan [12] improved traditional learning-to-rank using transfer learning, enhancing RankingSVM performance on P@N, MAP, and NDCG metrics. Tian [4] proposed a quality detection and ranking method that filters low-quality answers before ranking, achieving better accuracy. Zhou et al. [8] incorporated user profile information into SVMRank and ListNet, showing improved MRR and P@N performance when adding user-related features.

2.2.2 Building New Ranking Models Toba et al. [2] proposed a hybrid hierarchical classification model with six predefined question types, calculating category probabilities and answer quality in sub-models, achieving higher accuracy. Shen et al. [19] introduced a new architecture using similarity matrices with lexical and sequential information in deep structures, showing potential for

improving Q&A matching precision and outperforming baselines on DCG@P. Yuan et al. [5] proposed a hybrid community Q&A quality evaluation model outperforming PLSA and TSPR on NDCG@P. Zhao et al. [20] proposed a novel RNN-based heterogeneous asymmetric ranking model showing superior performance on NDCG, P@N, and Accuracy.

2.2.3 Building Respondent Ranking Models Some researchers propose ranking respondents first to identify experts, then using their answers as best answers. Liu et al. [13] proposed ZhihuRank, ranking user authority based on link structure and topic similarity between questions and experts, outperforming other algorithms on MRR and NDCG. Yang et al. [21] proposed CQARank, which finds experts with similar topic interests and high expertise based on Q&A voting history, outperforming others on MRR, P@N, and CDR@P. Liu et al. [11] further proposed RTEM (Related Topic Expertise Model), showing better performance than TEM on NDCG, Spearman, and Kendall metrics.

2.3 Research Review

Current research has limitations: (1) feature selection focuses mainly on answer and respondent features, with less consideration of voter features; (2) most studies emphasize algorithm optimization but lack comparative analysis of different learning-to-rank algorithm categories. This paper addresses these gaps by constructing a multidimensional feature system, comparing algorithm performance, analyzing feature contributions, and proposing optimization strategies.

3 Construction of Multidimensional Answer Ranking Feature System

3.1 Multidimensional Answer Features

Answer ranking aims to provide high-quality answers. Answer quality is the direct factor, while source reliability indirectly affects quality. Users are the core element, and voting mechanisms enable quality evaluation. We construct features from three dimensions: answer, respondent, and voter.

3.1.1 Answer Features We measure answer quality through four aspects: (1) **Length**: Longer answers contain more information and reflect greater effort [22,23]. (2) **Similarity to question**: Semantic similarity indicates quality—higher similarity means better alignment with information needs, while low similarity indicates useless answers that should be filtered [5]. (3) **Information entropy**: Reflects answer diversity; higher entropy indicates richer content [3]. (4) **Form**: Number of content types (text, images, links). External links, images, and charts support correctness and credibility, simplifying complex explanations. Rich answer forms are more likely to be high-quality.

3.1.2 Respondent Features Respondent features include: (1) **Expertise in the question domain**: Higher expertise indicates better answers [24]. We measure this through: (a) historical similar question answering expertise (a_{aSpecialty})—more similar answers and upvotes suggest domain expertise [5]; and (b) historical similar question asking expertise (a_{qSpecialty})—asking similar questions indicates prior knowledge. (2) **Community influence**: Influence stems from trustworthiness or expertise [25], measured through: upvotes received (a_{voteup}), follower count (a_{following}), answer count (a_{aNum}), and question count (a_{qNum}).

3.1.3 Voter Features Most communities rank answers by popular vote. Voter judgments reflect answer quality. Geerthik et al. [17] argued that answers with more influencer upvotes should rank higher. Voters also drive answer propagation, influencing others' perceptions. We consider only upvoter features (as downvote data is unavailable): average upvotes received by upvoters (v_{voteup}) and average follower count of upvoters (v_{following}). Higher-authority upvoters and lower-authority downvoters indicate better answers.

Table 1 shows the complete feature set and calculation methods.

3.2 Zhihu Community Feature System Analysis

Our experimental data comes from Zhihu, crawling 1,976 questions, 108,211 answers, 71,280 respondents, and 2,632,660 voters. After removing invalid data from deactivated accounts, we used 95,021 answers. We calculated features, examined correlations, and built the final 指标体系.

3.2.1 Feature Calculation Following Table 1 methods, we obtained feature values (Table 2).

3.2.2 Standardization and Correlation Analysis We applied min-max normalization to map data to [0,1]. Using Pearson correlation, we built the correlation matrix (Table 3). Correlation strength: 0.8-1 (very strong), 0.6-0.8 (strong), 0.4-0.6 (moderate), 0.2-0.4 (weak), 0-0.2 (very weak/none). We used 0.4 as the threshold.

3.2.3 Variable Selection Entropy (entropy) and length (length) correlated at 0.68—both measure information content, but entropy better captures information quality, so we removed length. a_{aSpecialty} correlated strongly with a_{aNum} (0.60) and a_{voteup} (0.96), so we kept the composite indicator a_{aSpecialty}. Similarly, a_{qSpecialty} correlated with a_{qNum} (0.67), so we kept a_{qSpecialty}. After removing length, a_{aNum}, a_{voteup}, and a_{qNum}, all remaining correlations were below 0.4.

Our final feature set contains eight features: question-answer similarity, answer entropy, answer form, respondent historical answering expertise, respondent historical asking expertise, respondent follower count, average upvoter upvotes, and average upvoter followers. Figure 1 [Figure 1: see original paper] shows the complete feature system, which integrates answer, respondent, and voter dimensions—unlike previous studies that used only one or two dimensions. We also applied Pearson correlation for feature selection, which most prior studies omitted, potentially affecting algorithm performance.

4 Experimental Analysis of Answer Ranking Based on Multidimensional Features

Our tasks: (1) compare algorithm performance, and (2) compare feature importance. Figure 2 [Figure 2: see original paper] shows the experimental flow.

4.1 Answer Ranking Experiment Based on Multidimensional Features

4.1.1 Experimental Tools Learning-to-rank datasets typically use binary relevance (1/0) or graded relevance (Perfect, Excellent, Good, Fair, Bad) [26], requiring costly manual annotation. Following Joachims et al. [27,28], we use upvote data as relevance signals. Voting resembles peer review and has proven effective [29]. Since Zhihu doesn't expose downvotes, we label answers with upvotes above the mean as 1, others as 0. This coarse-grained approach suits our data's relatively low discrimination. We labeled 30,261 instances as 1 and 64,760 as 0.

Table 4 lists the 11 ranking algorithms and tools used.

4.1.2 Evaluation Metrics We use NDCG@k and MRR. NDCG@k measures ranking quality for top-k results, considering both relevance degree and position. We evaluate at k=1,3,5,10. MRR measures the position of the first relevant answer—higher positions yield higher MRR. While NDCG@k evaluates the entire list, MRR focuses on the best answer. Using both provides comprehensive assessment.

4.2 Deep Learning Ranking Algorithm Example

This demonstrates the experimental workflow: dataset preparation, parameter tuning, and performance comparison. Proper parameter selection prevents overfitting/underfitting.

We set num_{features} to 8 (our feature count) and tuned num_{{{train}}_{{steps}}}. With initial value 20,000, evaluation metrics plateaued between 6,000-10,000 steps, degrading beyond 10,000 (Figure 3 [Figure 3: see original paper]). We tested values 6,000-10,000, selecting 8,000 as optimal (Table 6).

For num_{{{train}}_{{steps}}}=7,000, we performed 10-fold cross-validation.

Table 5 shows detailed metrics for each fold, with averages representing algorithm performance.

4.3 Results Analysis

4.3.1 Model Performance Analysis After 10-fold cross-validation and parameter tuning, we averaged results across folds (Table 7).

Algorithm Category Comparison: - **Neural network methods**: Pairwise-based RankNet (lr=0.00005) outperformed Listwise-based ListNet (lr=0.00005) on all metrics. - **Tree-based methods**: Pairwise-based MART (shrinkage/lr=0.05) slightly outperformed Listwise-based LambdaMART (shrinkage/lr=0.01). - **Boosting methods**: Pairwise-based RankBoost outperformed Listwise-based AdaRank (tolerance=0.002).

Listwise methods don't necessarily outperform Pairwise/Pointwise methods, possibly because Listwise struggles to find appropriate optimization targets and solvers [30].

Machine Learning Technology Comparison: - **Deep Learning** (num_{{{train}}_{{steps}}}=8,000) outperformed all others on NDCG@k and MRR. - **Tree-based methods** (Random Forest, MART, LambdaMART) outperformed traditional methods, with Random Forest (shrinkage/lr=0.1) performing best—likely because its ensemble of trees learning partial features reduces overfitting. - **Boosting (RankBoost), SVM (RankingSVM, c=0.01), Neural Network (RankNet, lr=0.00005), and Coordinate Ascent (tolerance=0.001)** outperformed AdaRank, ListNet, and Linear Regression (L2=1.0E-10).

4.3.2 Feature Importance Analysis We used Random Forest to assess feature importance through Gini importance—the frequency a feature is selected for splitting and its classification value. For binary classification at tree node τ with samples n and class counts n_k (k {0,1}), Gini impurity is i(τ)=1-p^2. Splitting by threshold t_ on variable creates child nodes τ_l and τr *with sample fractions p_l and p_r, reducing impurity by* $\Delta i(\tau)=i(\tau)$-p{li}(τ_l)-pri(τ_r). The importance of feature accumulates these reductions across all nodes and trees.

Figure 4 [Figure 4: see original paper] shows feature importance: 1. **Voter features** (average upvoter upvotes, average upvoter followers) scored highest—higher voter influence better reflects answer quality. 2. **Answer features**: Answer entropy and question-answer similarity scored high—content richness and semantic relevance are objective quality indicators. Answer form diversity was least important, showing users value content over presentation. 3. **Respondent features**: Historical answering expertise and influence (follower count) mattered more than historical asking expertise, as the former better indicates domain expertise.

5 Optimization Strategies for Social Q&A Community Answer Ranking

Based on our findings and community challenges, we propose strategies to increase ranking diversity and algorithm comprehensiveness.

5.1 Increasing Ranking Method Diversity

Our eight-feature system enables multiple ranking options: 1. **Rank by entropy or question-answer similarity**: Surfaces content-rich, semantically relevant answers first, filtering irrelevant or humorous answers and saving time. This also rewards dedicated non-expert respondents, enhancing community engagement. 2. **Rank by answer form** (text/images/links): Some users prefer concise visual explanations; others prefer credible answers with citations. Form-based ranking accommodates these preferences. 3. **Rank by respondent domain expertise**: Expert answers, though sometimes brief, efficiently solve problems. This benefits knowledge-seeking users. 4. **Rank by voter influence**: Users may value endorsements from influencers/experts more than popular votes.

Current ranking by upvotes/downvotes alone cannot meet diverse needs, as humorous answers often receive high votes but fail to satisfy knowledge needs.

5.2 Improving Algorithm Comprehensiveness

We extracted and analyzed features from three dimensions. Communities also possess clickstream data, access patterns, and backend logs for mining richer features. Our comparison shows deep learning outperforms other methods. Deep learning offers two advantages: 1. **Scalability**: Performance improves with sample size, suitable for big data era's full-sample analysis without validity/endogeneity tests, yielding more realistic results. 2. **Feature learning**: Learns relationships from low-dimensional dense features, reducing feature engineering needs compared to linear models.

Community developers should prioritize deep learning ranking algorithms.

Limitations and Future Work

Our study has limitations: incomplete dataset (no downvote data), insufficient sample size (only ~100K Zhihu answers), lack of cross-question-category analysis, and no consideration of herd effects where popular answers may receive more upvotes than professional ones. Future work will examine cross-category ranking differences and the relationship between upvotes and answer quality.

References

[1] Li Lei, He Daqing, Zhang Chengzhi. Review of Social Q&A Research [J]. Data Analysis and Knowledge Discovery, 2018, 2(7): 1-12. [2] Toba H, Ming Z

Y, Adriani M, et al. Discovering high quality answers in community question answering archives using a hierarchy of classifiers [J]. Information sciences, 2014, 47(8): 101-115. [3] Zhang Pengfei. Research on Question Retrieval and Answer Extraction Technologies for Online Q&A Communities [D]. Changsha: National University of Defense Technology, 2015. [4] Tian Zuohui. Answer Selection for Non-Factoid Questions [D]. Harbin: Harbin Institute of Technology, 2013. [5] Yuan Jian, Liu Yu. A Hybrid Community Q&A Answer Quality Evaluation Model [J]. Application Research of Computers, 2017, 34(6): 1708-1712. [6] Surdeanu M, Ciaramita M, Zaragoza H. Learning to rank answers to non-factoid questions from Web collections [J]. Computational linguistics, 2012, 37(2): 351-383. [7] Guo Shunli, Zhang Xiangxian, Tao Xing, et al. Research on Automated Evaluation of User-Generated Answer Quality in Social Q&A Communities—A Case Study of Zhihu [J]. Library and Information Service, 2019, 63(11): 118-130. [8] Zhou Z M, Lan M, Niu Z Y, et al. Exploiting user profile information for answer ranking in CQA [C]//21st World Wide Web conference 2012. Lyon: ACM Press, 2012: 767-774. [9] Cheng Yanan, Wang Yu. Research on Answer Ranking in Q&A Communities Based on Semantic and Sentiment Similarity [J]. Information Science, 2018, 36(8): 72-76, 83. [10] Lian Xin. Research on Several Key Issues in Community Q&A Systems [D]. Tianjin: Nankai University, 2014. [11] Liu Yu, Yuan Jian. A Q&A Community Candidate Answer Ranking Method Based on RTEM Model [J]. Electronic Science and Technology, 2016, 29(5): 130-134. [12] Yuan Liwei. Research on Transfer Learning Methods for Answer Ranking in Community Q&A Systems [D]. Kunming: Kunming University of Science and Technology, 2017. [13] Liu X, Ye S, Li X, et al. ZhihuRank: a topic-sensitive expert finding algorithm in community question answering Websites [C]//Advances in Web-based learning-ICWL 2015. Guangzhou: Springer International Publishing, 2015: 165-173. [14] Li B, King I, Lyu M R. Question routing in community question answering: putting category in its place [C]//ACM conference on information and knowledge management. Glasgow: ACM Press, 2011: 2041-2044. [15] Luo Yi, Cao Qian. Research on Answer Quality of Social Q&A Platforms Based on RIPA Method [J]. Library and Information Service, 2015, 59(3): 126-133, 25. [16] Yuan Yi, Yang Li. Analysis of User Behavior and Influencing Factors in Q&A Communities—A Case Study of Baidu Knows [J]. Library and Information Service, 2017, 61(22): 20-26. [17] Geerthik S, Rajiv G K, Venkatraman S. RespondRank: improving ranking of answers in community question answering [J]. International journal of electrical & computer engineering, 2016, 6(4): 1889-1896. [18] Cui Yujia, Zhang Yidi, Wang Peizhi, et al. Medical Data Feature Selection Algorithm Based on Multi-Evaluation Criteria Fusion [J]. Fudan Journal (Natural Science), 2019, 58(2): 250-255, 268. [19] Shen Y, Rong W, Sun Z, et al. Question/answer matching for CQA system via combining lexical and sequential information [C]//29th AAAI conference on artificial intelligence. Austin: AAAI Press, 2015: 275-281. [20] Zhao Z, Lu H, Zheng V W, et al. Community-based question answering via asymmetric multi-faceted ranking network learning [C]//Proceedings of the 31st AAAI conference on artificial intelligence. San Francisco: AAAI Press, 2017: 3532-3539. [21] Yang

L, Qiu M, Gottipati S, et al. CQARank: jointly model topics and expertise in community question answering [C]//ACM international conference on information & knowledge management. San Francisco: ACM Press, 2013: 99-108. [22] Jeon J, Croft W B, Lee J H, et al. A framework to predict the quality of answers with non-textual features [C]//The 29th annual international ACM SIGIR conference on research and development in information retrieval. Washington: ACM Press, 2006: 228-235. [23] Wang Le. Research on Knowledge Contribution and Interaction Quality in Social Q&A Communities [D]. Harbin: Harbin Institute of Technology, 2016. [24] Wang Xiuli. Research on the Influence Mechanism of Opinion Leaders in Online Communities—A Case Study of Social Q&A Community "Zhihu" [J]. Chinese Journal of International Press, 2014, 36(9): 47-57. [25] Perry-Smith J E, Mannucci P V. From creativity to innovation: the social network drivers of the four phases of the idea journey [J]. Academy of management review, 2017, 42(1): 53-79. [26] Liu T Y. Learning to rank for information retrieval [C]//International ACM SIGIR conference on research & development in information retrieval. Geneva: ACM Press. 2010: 1-112. [27] Radlinski F, Joachims T. Query chains: learning to rank from implicit feedback [C]//ACM SIGKDD international conference on knowledge discovery & data mining. Chicago: ACM Press, 2006: 217-226. [28] Radlinski F, Joachims T. Active exploration for learning rankings from clickthrough data [C]//ACM SIGKDD international conference on knowledge discovery & data mining. New York: ACM Press 2007: 570-579. [29] Hosseini M, Moore J, Almaliki M, et al. Wisdom of the crowd within enterprises: practices and challenges [J]. Computer networks, 2015, 17(15): 121-132. [30] Xiong Liyan, Chen Xiaoxia, Zhong Maosheng, et al. Review of PairWise Learning-to-Rank Algorithms [J]. Science Technology and Engineering, 2017, 17(21): 184-190. [31] Joachims T. Training linear SVMs in linear time [C]//ACM SIGKDD international conference on knowledge discovery & data mining. Philadelphia: ACM Press, 2005: 239-248. [32] Sourceforge [EB/OL]. [2020-05-08]. https://sourceforge.net/p/lemur/wiki/RankLib%20How%20to%20use/.

Author Contributions

Yi Ming: Research conceptualization and outline, final manuscript approval. Zhang Tingting: Data collection and analysis, initial draft writing. Li Ziqi: Data analysis optimization, manuscript revision.

Note: Figure translations are in progress. See original paper for figures.