Postprint of Answerer Discovery Research in Social Q&A Communities

Authors: Pan Mengya, Shen Wang, Dai Wang, Liu Jiayu

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

Abstract

[Purpose/Significance] Identifying professional answerers with high response likelihood in social Q&A communities can reduce waiting times for users seeking satisfactory answers, promote knowledge sharing among users, and facilitate the sustainable and healthy development of social Q&A communities. [Method/Process] Based on social capital theory and motivation theory, this study analyzes the motivations driving users to answer questions, proposes measurement indicators by integrating expert discovery research, and constructs a research model. Taking the Zhihu community as a case study, Python is employed to process experimental data, including feature extraction and labeling. The study utilizes three commonly used machine learning classification models—logistic regression, random forest, and XGBoost—for training and prediction. [Results/Conclusions] The effectiveness and superiority of the proposed method are validated through comparison with PageRank and HITS algorithms. This study provides valuable references for research on question routing, expert identification, and recommendation models for similar platforms such as health communities.

Full Text

Preamble

Volume 64, Issue 18, September 2020

Research on Respondent Discovery in Social Q&A Communities

Pan Mengya, Shen Wang, Dai Wang, Liu Jiayu

School of Management, Jilin University, Changchun 130022

Abstract: [Purpose/Significance] Identifying professional respondents with high answering probability in social Q&A communities can shorten the waiting time for questioners to obtain satisfactory answers, promote knowledge sharing

among users, and contribute to the sustainable and healthy development of social Q&A communities. [Method/Process] Based on social capital theory and motivation theory, this paper analyzes the motivations behind user answering behavior, proposes measurement indicators combined with expert discovery research, and constructs a research model. Taking Zhihu as a case study, Python was used for feature extraction and labeling of experimental data. Three common machine learning classification models—logistic regression, random forest, and XGBoost—were employed for training and prediction. [Result/Conclusion] Compared with PageRank and HITS algorithms, the effectiveness and superiority of the proposed method are verified. This research provides a reference for similar platforms, such as health communities, in question routing, expert identification, and recommendation model development.

Keywords: Social Q&A community; Expert discovery; Social capital theory; Motivation theory; Machine learning

Classification Numbers: G202, G206

DOI: 10.13266/j.issn.0252-3116.2020.18.009

The development of Internet technology has transformed how people search for and exchange information, leading to the rise and prosperity of online Q&A communities. These communities transcend temporal and spatial constraints, integrating user groups from diverse backgrounds and industries who share similar interests, goals, and practical experiences. They overcome the limitations of obtaining single pieces of existing information through search engines alone by transferring information, experience, and knowledge from users' minds into the community. Users can ask questions or answer inquiries across any domain and question type at any time, or engage in real-time communication with other community members through comments and private messages, sharing experiences and knowledge to solve practical problems.

However, in volunteer-based online Q&A communities, many user questions remain unanswered for extended periods or fail to receive professional, comprehensive, and satisfactory responses. Over time, questioners may become frustrated, potentially affecting the overall health of the community [1]. Therefore, identifying professional respondents with high answering probability for specific questions in social Q&A communities can help questioners obtain high-quality answers, reduce waiting times, and promote sustainable community development. Previous scholars have explored methods for identifying expert users in Q&A communities, but if these experts are constrained by various conditions and cannot respond promptly, the problem of unanswered questions persists. This study draws upon motivation theory and social capital theory, combined with expert discovery research, to identify professional respondents with high answering probability to address these challenges. Multiple methods are employed to validate the research model and identify the optimal algorithm, differing from previous studies that relied on single algorithms. Experimental results confirm

the effectiveness and superiority of this approach.

This research focuses on the popular social Q&A community Zhihu as a case study (as of January 2019, Zhihu ranked 3rd among social networking sites in China and 90th globally according to Alexa, with approximately 5 million daily IP visits). Data was collected from Zhihu's medical topic area, as medicine represents a domain where both expert and ordinary users can actively participate, making the sample highly representative.

2. Literature Review

The objective of this study is to identify professional respondents with high answering probability in social Q&A communities. This section reviews relevant research on expert discovery and user knowledge sharing.

2.1 Expert Discovery

Expert discovery in social Q&A communities involves identifying users who possess professional knowledge and authority from among numerous respondents [2]. Scholars have employed various research methods and perspectives to investigate expert discovery in Q&A communities:

(1) **Content-based approaches**: J. Weng et al. [3] utilized TwitterRank algorithm for topic-sensitive expert ranking based on users' tweet distributions and content homogeneity. A. Pal et al. [4] identified topic authorities through content clustering using Gaussian mixture models. Z. Yan et al. [5] employed tensor and topic models to study latent semantic relationships between questions and answerers, optimizing potential respondent ranking through AUC maximization.

(2) **User feedback behavior analysis**: X. Cheng et al. [6] used user feedback as relevant label words to build topic models, combining user expertise features for expert discovery. J. Shen et al. [7] recommended experts using weighted HITS algorithm based on likes, comments, and best answer selections. S. Patil et al. [8] distinguished expert from non-expert behaviors, identifying experts using statistical models based on four indicators: user activity features, answer quality features, linguistic features, and temporal features.

(3) **Social network relationships**: Gong Kaile et al. [9] expanded user modeling based on "question-user" propagation networks, identifying experts through answer quality weighting. S. Yarosh et al. [10] constructed "task-topic" cross-scenarios using experts' social relationships and professional knowledge, selecting experts from recommendation lists via the SmallBlueFind system. S. Ghosh et al. [11] mined Twitter list metadata to identify topic experts using the Cognos system.

(4) **Authority and reputation modeling**: D. R. Liu et al. [12] modeled users through linear combinations of topic preference, reputation, and au-

thority, where topic preference was calculated via text similarity between expert profiles and target questions, reputation derived from historical answer counts and best answers, and authority obtained through link analysis algorithms. L. Hong et al. [13] embedded probabilistic latent semantic analysis into user reputation modeling for expert discovery using PageRank. Lin Hongfei et al. [14] proposed an expert discovery method based on user category participation, calculating expert scores and participation scores for each category using PageRank and HITS.

In summary, previous research on expert discovery has utilized various measurement indicators and techniques, but most studies relied on text similarity or auxiliary local features, employed single technical methods for model validation, and focused primarily on technical approaches with limited theoretical grounding. This study aims not only to identify experts but also to find experts with high answering probability. Based on social capital theory and motivation theory, it investigates user answering motivations from multiple features and seeks suitable technical methods across multiple machine learning models, thereby enriching prior expert discovery research.

2.2 Knowledge Sharing

Knowledge sharing refers to the behavior of expressing and disseminating knowledge through various media [15]. Answering questions in social Q&A communities constitutes knowledge sharing. Academic research on user knowledge sharing behavior has been extensive, primarily theoretical. This study adopts the well-established motivation theory and social capital theory as its theoretical foundation.

Motivation Theory posits that human behavior is motivation-driven, with motivation being a necessary prerequisite for knowledge sharing [16]. Scholars categorize user motivations in virtual communities into intrinsic motivations (e.g., personal interest [17], altruism, desire for recognition [18-20]) and extrinsic motivations (e.g., reputation [18-21], benefits [17], external rewards [19-20,22], acquiring useful information and expertise [23]). These motivations enable users to obtain tangible or intangible benefits or achieve self-fulfillment, significantly influencing knowledge sharing behavior. This study analyzes user answering probability from internal motivations (need satisfaction, altruism) and external motivations (time and benefits).

Social Capital Theory suggests that social capital—the relational networks and embedded resources possessed by individuals or social networks—strongly influences the extent of knowledge sharing. The theory comprises three dimensions: structural, relational, and cognitive [24]. Studies by Zhao Ling et al. [25], C. M. Chiu et al. [26], L. Zhao et al. [27], H. H. Chang et al. [19], and H. F. Lin [28] demonstrate that community interaction relationships, such as trust and reciprocity, affect users' sense of belonging (membership) in virtual communities, thereby influencing knowledge sharing activity. B. Vanden Hoof et al. [29]

argue that community trust, identification, and users' knowledge sharing capabilities and willingness are key factors affecting knowledge sharing. This study analyzes user answering potential from the three dimensions of social capital theory, focusing on reciprocity, common language among members, and users' social relationship networks.

3. Feature Indicators for Respondent Discovery in Social Q&A Communities

3.1 Analysis of User Answering Motivations

3.1.1 Motivation Theory Perspective From the motivation theory perspective, user answering behavior in social Q&A communities is driven by the following factors (see Figure 1 [Figure 1: see original paper]):

Internal motivations include need satisfaction and altruism. Need satisfaction refers to users answering questions related to their interests for self-fulfillment, or improving their knowledge structure and self-development by sponsoring paid live sessions (Zhihu Live is a real-time Q&A feature where users can participate live or watch replays, rating sessions by awarding stars. Users can become hosts, sharing experiences via audio, images, video, or text, with free or paid admission). Altruistic motivations drive users to answer questions broadly or host free/ultra-low-cost Zhihu Live sessions to exchange information and share experiences.

External motivations primarily involve time availability and material benefits. Time is a prerequisite for community activity levels, while monetary benefits and reputation incentivize high-quality content creation. Monetary motivation manifests in hosting paid live sessions for knowledge monetization, while reputation motivation is evident in the behavior patterns of verified institutional and individual users (frequently mentioning certification-related information).

3.1.2 Social Capital Theory Perspective From the social capital theory perspective, user answering motivations are analyzed across three dimensions: relational, cognitive, and structural (see Figure 2 [Figure 2: see original paper]):

Relational social capital manifests as reciprocal relationships among community members, including intangible resource reciprocity (e.g., user D liking or thanking user C's content) and tangible resource reciprocity (e.g., user D hosting paid live sessions for knowledge monetization while user F pays to acquire knowledge).

Cognitive social capital refers to shared vision and language among community members, where Zhihu users gather under specific topics due to shared interests, hobbies, or language, striving to acquire and share knowledge to enrich domain expertise.

Structural social capital concerns users' positions in social network struc-

tures, primarily related to follower counts and the importance of nodes in social networks.

3.2 Expert Discovery Feature Indicators

Based on previous expert discovery research, the following indicators measure user expertise in social Q&A communities:

1. **User Credibility**: In social Q&A communities, users are both audience members and content creators. Information credibility is closely tied to publishers, representing users' subjective trust in information beyond simple truthfulness [30]. This study evaluates credibility through user profile information and community interaction behaviors.
2. **User Professionalism**: Experts are users who have answered similar questions [2]. This study measures professionalism through historical answer topic distribution and content quality, including detail, clarity, professionalism, presence of topic-related keywords, use of charts/links for supplementation, and answer length.
3. **User Authority**: Users tend to socialize online with similar others [28], forming "circles" based on interests and knowledge domains [31]. Zhihu users develop "following," "followed by," and "follower" relationships (see Figure 3 [Figure 3: see original paper]), creating a vast, clear social network. Authority values derive from these networks via HITS or PageRank algorithms [12].

3.3 Respondent Feature Indicators and Measurement

Based on the above motivation analysis and previous expert discovery research, this study proposes feature indicators (see Figure 4 [Figure 4: see original paper]) and measurement methods.

3.3.1 User Credibility User credibility measurement ensures answer credibility. User profiles include nickname, avatar, gender, location, industry, career experience, education, and personal introduction. Users may register with real names or anonymously. Profile completeness reflects credibility to some extent. This study measures completeness by counting non-empty fields among six items: gender, location, industry, career experience, education, and personal introduction. Verified users (individual or institutional) have higher credibility than unverified or anonymous users.

Interaction behavior data validates whether users are bots. Beyond Q&A and follow relationships, interactions are measured through likes received, thanks received, and favorites received. The TOPSIS method normalizes these metrics to extract feature values while minimizing information loss.

3.3.2 User Activity Active users are more likely to answer new questions. Activity is measured through five indicators: historical answer count, question count, article count, live sessions hosted, and public edit participation count in 2018. All are benefit-type indicators processed via TOPSIS to obtain activity level (C-value).

3.3.3 User Professionalism Most Zhihu users don't explicitly state their interest topics. Describing topic distribution helps identify user interests. When user topics align with question topics, users are more likely to answer due to internal need satisfaction motivations. Professionalism is measured through:

(1) **Topic similarity**: Matching question keywords with user interest keywords, including similarity between user profile information and new questions, and between user topics and question topics. Information points are vectorized to calculate cosine distance. The cosine similarity between new question vector H and user topic vector V is:

$$Sim(\vec{H}, \vec{V}) = \frac{\vec{H} \cdot \vec{V}}{\|H\| \cdot \|V\|}$$

where H and V are n-dimensional vectors. Values closer to 1 indicate greater similarity.

Text preprocessing involves Chinese word segmentation using Jieba, filtering stop words, punctuation, special symbols, and internet slang. Each user's profile or answer is treated as a document, with all user documents forming the corpus for tf-idf calculation (see Figure 5 [Figure 5: see original paper]).

LDA topic model extracts key topic features from text. LDA is a three-layer Bayesian probability model (word-topic-document) where both document-to-topic and topic-to-word distributions follow multinomial distributions. Using the bag-of-words approach, LDA simplifies complex problems by representing documents as word frequency vectors.

(2) **Content quality**: Includes detail, clarity, and professionalism, measured by image/hyperlink count, likes received, thanks received, and answer length. Images and links are counted via html tags (and). Most users average fewer than 0.3 images/links per answer.

3.3.4 User Authority Authority is measured through social network importance and content quality impact. Higher authority attracts more followers, increasing answer visibility, likes, and shares, creating a virtuous cycle.

(1) **Content impact**: Measured by likes, shares, and favorites. High-like answers are prioritized, enhancing user influence. Since shares and favorites aren't displayed per answer, total likes per user are used.

(2) **PeopleRank algorithm**: Similar to PageRank, it treats the social network as a directed graph with users as nodes and follow relationships as edges (see Figure 6 [Figure 6: see original paper]). For any user A:

$$PR(A) = (1-d) + d \sum_{p_i \in F(A)} \frac{PR(p_i)}{C(p_i)}$$

where p_i represents users, $C(p_i)$ is the number of users followed by p_i, and d is the damping coefficient representing the probability that follow relationships may change authority levels. Initial PR values are assigned and iteratively calculated until convergence.

3.3.5 Other Feature Indicators Features requiring no calculation include: institutional user status (motivated by reputation and status), live session count and price (self-improvement through sponsored sessions, knowledge monetization through high-priced sessions, altruism through free/ultra-low-cost sessions), and answer count (altruism manifestation). These are measured through simple statistical methods.

4. Respondent Discovery Process and Algorithm

4.1 Discovery Process

Python scripts collected data from Zhihu, followed by cleaning and preprocessing. Feature extraction included: (1) profile completeness; (2) user activity; (3) community interaction; (4) historical content features; (5) social network importance; (6) question-topic similarity. The dataset was split 60% for training and 40% for testing. Three machine learning models (logistic regression, random forest, XGBoost) were used to build optimal binary classification models to predict answering probability for unanswered questions (see Figure 7 [Figure 7: see original paper]).

4.2 Discovery Algorithm

The algorithm processes extracted features using three models. Pseudocode is described in Table 1 .

5. Experimental Construction and Comparative Analysis

5.1 Dataset and Preprocessing

5.1.1 Experimental Dataset Data was collected from Zhihu's medical topic area using Python web scraping from January 1, 2018, to April 30, 2019. The dataset includes: (1) user identity information (ID, industry, education, profile); (2) verification status; (3) historical Q&A data; (4) articles; (5) live sessions; (6) follower/fan IDs. The dataset contains 318 users, 844 questions, 65,352 answers, 31,243 follow relationships, and 276,379 fan relationships.

5.1.2 Data Preprocessing Twelve anonymous users with missing profile data were removed, leaving 306 valid users with 65,251 historical answers, 31,243 follow relationships, and 276,379 fan relationships. Preprocessing included: (1) SQL-based deduplication; (2) Python-based text cleaning (removing html tags, word segmentation, stop word removal using Jieba and Harbin Institute of Technology stop word list, supplemented with custom stop words).

5.2 Feature Extraction and Analysis

5.2.1 User Credibility **Profile completeness**: Most users have completeness below 40%, with few above 60%. Only 8 users (2.6%) are verified.

Interaction behavior: Calculated via TOPSIS based on likes, thanks, and favorites (see Table 2). After normalization, optimal and worst values (Z^+ and Z^-) are identified. For example, for Evaluation Object 1:

$$Z^+ = (0.9578, 0.9835, 0.7258)$$

$$Z^- = (0, 0, 0)$$

Resulting in $D^+ \approx 1.552389$, $D^- \approx 0.000487$, and $C_1 \approx 0.0003135$ (see Table 3).

5.2.2 User Activity Activity level is measured through answer count, question count, article count, live sessions, and public edit participation. TOPSIS results for 10 users are shown in Table 4 .

5.2.3 User Professionalism **Topic similarity**: LDA extracts 10 topics per user (8 words each). Question topics are categorized into 10 hot topics: "medical," "gaming," "technology," "film," "food," "life," "employment," "education," "intimate relationships," and "income." Topic dictionaries are constructed, and similarity between user topics and question topics is calculated as a feature value. Users are labeled 0/1 based on historical answers to topic-related questions (see Table 5).

Content quality: Most users produce ~100 answers annually, with few exceeding 3,000. Average answer length is ~70 words, with detailed answers exceeding 100 words. The like-to-answer ratio (γ) shows 83% of users average fewer than 10 likes per answer, 14.05% average 10-100 likes, 1% exceed 1,000 likes, and 2.6% are verified institutional users (see Figure 8 [Figure 9: see original paper]).

5.2.4 User Authority Based on 31,243 follow and 276,379 fan relationships, PeopleRank calculates user importance in the social network (see Figure 9 [Figure 9: see original paper]). Larger nodes indicate higher importance and authority.

5.2.5 Other Feature Values Institutional users comprise 0.6% of the dataset. Regarding Zhihu Live: 73.4% of users sponsor sessions for self-improvement (some spending over 1,000 yuan), while others host sessions for knowledge monetization (average price >25 yuan) or altruism (free/ultra-low-cost, >2 hours).

5.3 Experimental Results and Comparative Analysis

The dataset was split 60% training and 40% testing. Three machine learning models were trained and evaluated (see Table 6):

- **Logistic Regression**: accuracy 59.8%, f1-score 57.0%, roc_{auc} 66.1%
- **Random Forest**: accuracy 79.1%, f1-score 70.2%, roc_{auc} 75.2%
- **XGBoost**: accuracy 86.4%, f1-score 79.7%, roc_{auc} 82.4%

XGBoost achieved the best performance (86.4% accuracy, 79.7% f1-score). Its PR and ROC curves are shown in Figure 10 [Figure 10: see original paper].

The prediction score is calculated as:

$$Score = AL \cdot (\alpha_1 feature_1 + \alpha_2 feature_2 + \alpha_3 feature_3 + ... + \alpha_n feature_n)$$

Sample prediction results are shown in Table 7 .

Comparison with PageRank and HITS algorithms (see Table 8) demonstrates the proposed model's superiority, as traditional methods ignore user behavior information and topic preferences.

6. Conclusion and Future Directions

This study collected, cleaned, and analyzed one year (2018) of data from Zhihu's medical topic. Based on social capital and motivation theories, feature indicators and a research model were constructed. Python was used to convert data into model features, with 0/1 labels assigned based on historical topic-related answers. Using 60% training and 40% testing data, logistic regression, random forest, and XGBoost models were trained. XGBoost achieved the highest accuracy (~86%), demonstrating good performance.

Limitations include: (1) exclusion of anonymous users who sometimes provide high-quality answers; (2) objective measurement of subjective answering motivations. Future research should address these issues.

References

[1] Le L T, Shah C. Retrieving people: identifying potential answerers in community question-answering [J]. Journal of the Association for Information Science and Technology, 2018, 69(10): 1246-1258.

[2] Liu X, Croft W B, Koll M, et al. Finding experts in community-based question-answering services [C]//Proceedings of the 14th ACM international conference on information and knowledge management. New York: Association for Computing Machinery, 2005: 315-316.

[3] Weng J, Lim E P, Jiang J, et al. Twitterrank: finding topic-sensitive influential tweeters [C]//Proceedings of the third ACM international conference on web search and data mining. New York: Association for Computing Machinery, 2010: 261-270.

[4] Pal A, Counts S. Identifying topical authorities in microblogs [C]//Proceedings of the fourth ACM international conference on web search and data mining. New York: Association for Computing Machinery, 2011: 45-54.

[5] Yan Z, Zhou J. Optimal answerer ranking for new questions in community question answering [J]. Information processing & management, 2015, 51(1): 163-178.

[6] Cheng X, Zhu S, Chen G, et al. Exploiting user feedback for expert finding in community question answering [C]//Proceedings of the 2015 IEEE international conference on data mining workshop. Washington, D.C.: IEEE Computer Society, 2015: 295-304.

[7] Shen J, Shen W, Fan X, et al. Recommending experts in Q&A communities by weighted HITS algorithm [C]//2009 international forum on information technology and applications. New York: Institute of Electrical and Electronics Engineers, 2009: 151-154.

[8] Patil S, Lee K. Detecting experts on quora: by their activity, quality of answers, linguistic characteristics and temporal behaviors [J]. Social network analysis and mining, 2016, 6(1): 1-11.

[9] Gong Kaile, Cheng Ying. Research on expert discovery methods in network Q&A communities based on "question-user" networks [J]. Library and Information Service, 2016, 60(24): 115-121.

[10] Yarosh S, Matthews T, Zhou M. Asking the right person: supporting expertise selection in the enterprise [C]//Proceedings of the SIGCHI conference on human factors in computing systems. Austin: Association for Computing Machinery, 2012: 2247-2256.

[11] Ghosh S, Sharma N, Benevenuto F, et al. Cognos: crowdsourcing search for topic experts in microblogs [C]//Proceedings of the 35th international ACM SIGIR conference on research and development in information retrieval. New York: Association for Computing Machinery, 2012: 575-590.

[12] Liu D R, Chen Y H, Kao W C, et al. Integrating expert profile, reputation and link analysis for expert finding in question-answering websites [J]. Information processing & management, 2013, 49(1): 312-329.

[13] Hong L, Yang Z, Davison B D, et al. Incorporating participant reputation

in community-driven question answering systems [C]//2009 international conference on computational science and engineering-volume 04. Washington, D.C.: IEEE Computer Society, 2009: 475-480.

[14] Lin Hongfei, Wang Jian, Xiong Daping, et al. Community Q&A expert discovery method based on category participation [J]. Computer Engineering and Design, 2014, 35(1): 333-338.

[15] Sharratt M, Usoro A. Understanding knowledge-sharing in online communities of practice: motivators, barriers, and enablers [J]. Advances in developing human resources, 2008, 10(4): 541-554.

[16] Ardichvili A. Learning and knowledge sharing in virtual communities of practice: motivators, barriers, and enablers [J]. Advances in developing human resources, 2008, 10(4): 541-554.

[17] Razmerita L, Kirchner K, Nielsen P. What factors influence knowledge sharing in organizations? A social dilemma perspective [J]. Journal of knowledge management, 2016, 20(6): 1225-1246.

[18] Huang W, Zhao P. Research on influencing factors of user knowledge sharing behavior in virtual communities [J]. Information Science, 2016, 34(4): 68-73, 103.

[19] Chang H H, Chuang S S. Social capital and individual motivations on knowledge sharing: participant involvement as a moderator [J]. Information & management, 2011, 48(1): 9-18.

[20] Cho H, Chen M H, Chung S. Testing an integrative theoretical model of knowledge-sharing behavior in the context of Wikipedia [J]. Journal of the American Society for Information Science and Technology, 2010, 61(6): 1198-1212.

[21] Zhang Y, Fang Y, Wei K K, et al. Exploring the role of psychological safety in promoting the intention to continue sharing knowledge in virtual communities [J]. International journal of information management, 2013, 33(3): 539-552.

[22] Wierenga C, De Ruyter K. Beyond the call of duty: why customers contribute to firm-hosted commercial online communities [J]. Organization studies, 2007, 28(3): 347-376.

[23] Butler B, Sproull L, Kiesler S, et al. Community effort in online groups: who does the work and why [J]. Leadership at a distance: research in technologically supported work, 2002, 54(1): 171-194.

[24] Nahapiet J, Ghoshal S. Social capital, intellectual capital, and the organizational advantage [J]. Academy of management review, 1998, 23(2): 242-266.

[25] Zhao L, Lu Y, Wang B, et al. Cultivating the sense of belonging and motivating user participation in virtual communities: A social capital perspective [J]. International journal of information management, 2012, 32(6): 574-588.

[26] Chiu C M, Cheng H L, Huang H Y, et al. Exploring individuals' subjective well-being and loyalty toward social networking sites from the perspective of network externalities: the Facebook case [J]. International journal of information management, 2013, 33(3): 539-552.

[27] Zhao L, Lu Y, Wang B, et al. Cultivating the sense of belonging and motivating user participation in virtual communities: A social capital perspective [J]. International journal of information management, 2012, 32(6): 574-588.

[28] Lin H F. Determinants of successful virtual communities: contributions from system characteristics and social factors [J]. Information & management, 2008, 45(8): 522-527.

[29] Van Den Hoof B, Elving W, Meeuwsen J M, et al. Knowledge sharing in knowledge communities [C]//Huy Smith M, Wenger E, Wulf V. Communities and technologies • January 2003. Netherlands: Kluwer, B.V., 2003: 119-141.

[30] Jiang Shengyi, Chen Dongyi, Pang Guansong, et al. Review of microblog information credibility analysis [J]. Library and Information Service, 2013, 57(12): 136-142.

[31] Yang C, Ding H, Yang J, et al. Research of microblog community detection based on clustering analysis [J]. Advances in information sciences and service sciences, 2013, 5(3): 25-31.

Author Contributions

Pan Mengya: Paper writing and revision, data processing and analysis;
Shen Wang: Research design, final version revision;
Dai Wang: Data analysis;
Liu Jiayu: Paper revision.

Social Question Answering Community Respondent Discovery Research

Pan Mengya, Shen Wang, Dai Wang, Liu Jiayu

School of Management, Jilin University, Changchun 130022

Abstract: [Purpose/Significance] Identifying professional answerers with high probability in social Q&A communities can shorten waiting time for satisfactory answers, promote knowledge sharing, and support sustainable community development. [Method/Process] Based on social capital and motivation theories, this paper analyzes user answering motivations, proposes measurement indicators combined with expert discovery research, and constructs a research model. Using Zhihu as a case study, Python was employed for feature extraction and labeling. Three machine learning classification models—logistic regression, random forest, and XGBoost—were used for training and prediction. [Result/Conclusion] Compared with PageRank and HITS algorithms, the method's

effectiveness and superiority are verified. This research provides reference for similar platforms in question routing, expert identification, and recommendation models.

Keywords: Social Q&A community; Expert discovery; Social capital theory; Motivation theory; Machine learning

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