IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 9, 2020, accepted March 25, 2020, date of publication March 30, 2020, date of current version April 14, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2984342

Helping the Ineloquent Farmers: Finding Experts for Questions With Limited Text in Agricultural Q&A Communities

XIAOXUE SHEN[1], ADELE LU JIA[1], SIQI SHEN[2], AND YONG DOU[2]
[1]College of Information and Electrical Engineering, China Agricultural University, Beijing 100089, China
[2]School of Computer, National University of Defense Technology, Changsha 410073, China

Corresponding authors: Adele Lu Jia (ljia@cau.edu.cn) and Siqi Shen (shensiqi@nudt.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61502500 and Grant 61602500, and in part by the Key Program of National Natural Science Foundation under Grant 61732018.

ABSTRACT Nowadays, hundreds of thousands of farmers in China seek online in agricultural Q&A communities, such as Farm-Doctor, for agricultural advice. As in many other Q&A communities, the key design issue is to find experts to provide timely and suitable answers. State-of-the-art approaches often rely on extracting topics from the question texts, however, the major challenge here is that questions in agricultural Q&A communities often contain limited textual information. To solve this problem, in this article, we conduct an extensive measurement on Farm-Doctor, which consists of over 690 thousand questions and over 3 million answers, and we model Farm-Doctor as a heterogeneous information network that incorporates rich side information. We propose a novel approach based on graph neural network to accurately recommend for each question the users that are highly likely to answer it. With an average income of fewer than 6 dollars a day, our method helps these less eloquent farmers with their cultivation and hopefully provides a way to improve their lives.

INDEX TERMS Question and answering, question routing, network representation learning.

I. INTRODUCTION

Community-based Question and Answering (CQA) systems have become popular knowledge-sharing platforms where users get answers to the questions they raised. They have received great attention both in industry and in academia [1], [2]. One of the most important goals of CQA systems is to provide an asker with a suitable answer in the shortest possible time, i.e., the so-called *question routing* problem. In contrast to previous works [2]–[4] that focus on general CQA websites such as Quora and Yahoo! Answers, in this work, we analyze the question routing problem in a CQA platform named Farm-Doctor that is exclusive for agricultural knowledge.

Though with limited income (on average less than 6 dollars a day), nowadays in China, hundreds of thousands of farmers managed to connect to the internet and seek online in agricultural Q&A communities like *Farm-Doctor* for cultivation advice. Accurate *question routing*, i.e., finding experts to answer the questions, will provide timely advice for their cultivation and potentially improve their lives. However, question routing in Farm-Doctor faces a major challenge, i.e., farmers are rather ineloquent, and questions are often raised with very *limited textual information*.

In most Q&A communities, questions are described in natural languages, and question routing is often performed through extracting topics from the rich textual information [2], [4]. In contrast, as depicted in Figure 2, farmers in Farm-Doctor raise their questions mostly through pictures, along with simple questions like *"which is the problem? ař* and *"what should I do?"* On the other hand, although image recognition has received great attention and success both in industry and in academia, efficient tools on identifying crop diseases and even crops are still missing. As a consequence, it is difficult to infer the topics from the questions (texts or images) alone.

In this work, we conduct, to the best of our knowledge, the first analysis of the question routing problem without using textual information. To this end, we have obtained the whole knowledge repository of Farm-Doctor, which

The associate editor coordinating the review of this manuscript and approving it for publication was Hong-Mei Zhang.

contains over 600 thousand questions and over 3 million answers. Based on user activities we model Farm-Doctor as a heterogeneous information network (HIN) that encodes multiple types of entities (users, crops, questions, answers) and relationships. And we propose a novel graph neural network (GNN) model to learn, through a deep learning approach, the low-dimensional vector embedding of the entities. Then the embedding vectors are used as input features for machine-learned classifiers. In this way, without using any textual information, our method is able to recommend timely and accurately, for the newly posted questions, the potential users that most likely will answer the questions.

Our analysis of Farm-Doctor mainly consists of three parts:

Measuring Farm-Doctor. In this work, we have obtained the whole knowledge base of Farm-Doctor. Our dataset covers all the 697,695 questions raised before April 21st, 2018, along with the information on the associated 3,179,333 answers, 438 crops, and 305,359 users. The information we obtained includes not only the basic question characteristics but also user activities such as who raised/answered question at what time, which crop is tagged in which question and is interested to which user. The dataset is publicly available through requests to the first author.

Characterizing question and answer dynamics. We first provide an analysis of the scale and the characteristics of the question repository in Farm-Doctor. We examine the number of answers received by each question and the timeliness of the answers. We find that questions in Farm-Doctor normally attract a few answers shortly after they are raised, but as time passes by, they no longer receive any attention. Then, we analyze the user activities in terms of the number of questions they raised and answered. We find a highly skewed activity level of the users, with a small number of users raising and answering a large number of questions. These results all indicate the need for a proper question routing method in Farm-Doctor, so that questions will get more answers and hopefully that less active users will be encouraged to answer the questions personalized recommended to them.

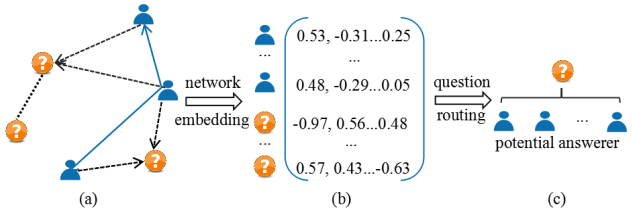

FIGURE 1. The overview of the question routing.

Question routing. Figure 1 shows an overview of question routing. To tackle the limited textual information problem in Farm-Doctor, we build a HIN model based on a variety of relationships (Figure 1 (a)) and we propose a heterogeneous GNN model to learn the low-dimensional embedding of the questions and the users (Figure 1 (b)). Taking the learned representations as input features, we build machine-learned classifiers to recommend timely and accurately, for the newly posted questions, the potential users that most likely will answer the questions (Figure 1 (c)).

The main contributions of this work are as follows:

- We obtain the whole knowledge repository of Farm-Doctor (until April 2018) that contains information of 305,359 users, 697,695 questions, and 3,179,333 answers (Section 2). Our dataset is publicly available upon request to the first author.
- We analyze the basic characteristics of the question repository and the user activities in Farm-Doctor (Section 2).
- We propose a heterogeneous graph model that incorporates a variety of relationships among users, questions and crops. Based on this model, we propose a novel GNN method to, for each question, accurately predict the potential answerers (Section 4).

II. METHODOLOGY AND THE FARM-DOCTOR DATASET

In this section, we first give a brief introduction to Farm Doctor. Then, we introduce the dataset used throughout this article. Finally, we reveal the basic characteristics of Farm-Doctor. This analysis will provide insights into the question routing problem analyzed later in Section 4.

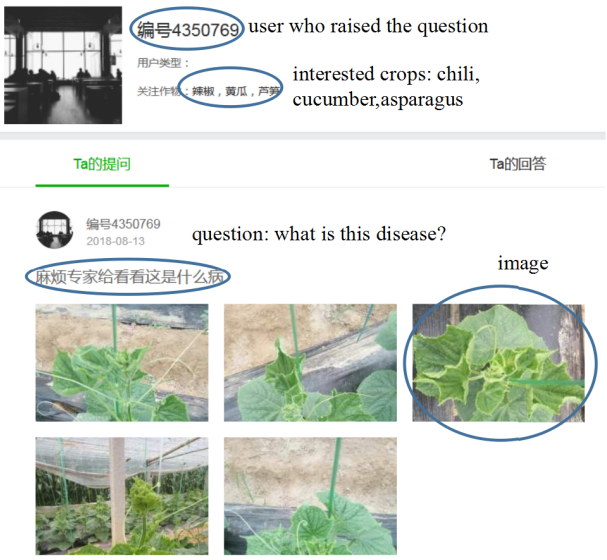

FIGURE 2. An example of a question raised in Farm-Doctor.

A. AN OVERVIEW OF FARM-DOCTOR

Farm-Doctor is a CQA platform exclusive for agricultural questions. It is one of the largest CQA platforms in China that provide farmers with advice on cultivation, such as crop disease detection and treatment recommendation. As in general CQA platforms, users in Farm-Doctor can raise and answer questions, vote to the answers, and follow each other. In addition, they can specify the crops that are interested in them and tag their questions with related crops. Figure 2 shows an example of the questions in Farm-Doctor. We can

see that, different from general CQA platforms, questions in Farm-Doctor contain very limited textual information.

B. DATASET

Farm-Doctor identifies each of its users with a unique numerical number in the increasing order. Each identifier corresponds to a webpage with detailed user information that can be obtained with crawlers. We have obtained the whole knowledge base (until April 2018) of Farm-Doctor with detailed information of 305,359 users, 697,695 questions, and 3,179,333 answers.

For each user, we obtain the list of questions that he has asked and answered, and the crops he follows. For each question, we obtain its description, the time when it was raised, the related crops in its tags, and the list of its answers. For each answer, we obtain the time when it was left, the identity of the answer, and the tags for the related crop diseases. In total, 73,463 users have raised at least one question (named *askers*) and 43,323 users have answered at least one question (named *answerers*). Table 1 shows the basic statistics of our dataset.

TABLE 1. Basic statistics of the farm-doctor dataset.

# questions	697,695
# answers	3,179,333
# answers per question	5
# users	305,359
# askers	73,463
# answerers	43,323
# questions per asker	9
# answers per answerer	73
# crops	438

C. CHARACTERISTICS OF FARM-DOCTOR

1) QUESTIONS

We first analyze the basic characteristics of all the 697,695 questions raised in the Farm-Doctor by the end of April 2018. We investigate the number of answers they received and the timeliness of these answers.

a: NUMBER OF ANSWERS RECEIVED

Figure 3 plots the Cumulative Distribution Function (CDF) of the number of answers received by each question. In total, 7.7% (54,084) questions have received more than 10 answers while 65.4% (456,294) questions have received fewer than 5 answers. We also observe that 18.3% (127,678) questions have received only one answer and 4.4% (30,861) questions have not been answered at all.

b: TIMELINESS OF THE ANSWERS

We analyze the timeliness of the answers by examining the delay of the answers. In Figure 4 we show the CDFs of the delay of the first answer and the time span of all the answers received by each question, respectively. We find that 94.5% (630,158) questions receive their first answers on the same day when the questions are raised, and that 76.5% (510,128) questions receive all their answers within one day.

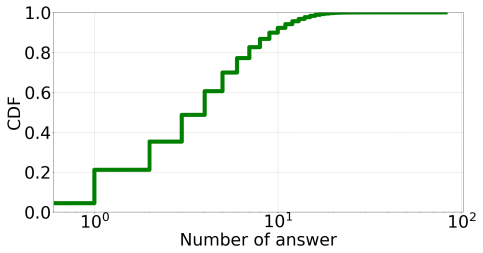

FIGURE 3. CDF of the number of answers received by each question.

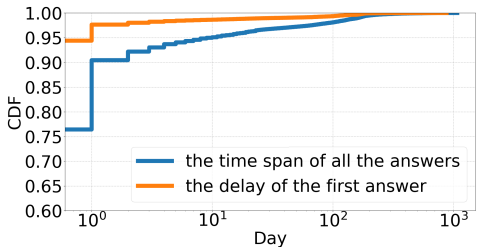

FIGURE 4. CDFs of the delay of the first answer and the time span of all the answers received by each question.

c: DISCUSSION

From the above results, we conjecture that, due to the absence of a proper question routing algorithm, questions in Farm-Doctor can only attract some answers when they are first raised and as time passes by they no longer receive any attention. To test our intuition, we have calculated the Spearman Ranking Correlation Coefficient (SRCC)[1] between the number of answers a question received and the time when the question was raised, and we do not find strong correlations between these two metrics (with a SRCC of 0.2442). In the agricultural area, users seek to get timely answers, preferably multiple opinions, to solve crop diseases and to reduce economic losses. The above results indicate the need for a better question routing method in Farm-Doctor that will route the questions to potential answerers accurately and timely.

2) USERS

For any CQA platform, the user activity level provides important knowledge for community development and prosperity. In this section, we analyze the user activities of the 305,359 users in Farm-Doctor.

a: RAISING AND ANSWERING QUESTIONS

We show in Figure. 5 the CDFs of the number of questions raised and answered by each user, respectively. We find the user activity level is highly skewed for both raising and answering questions. More specifically, while 5% (28,132) users have raised more than 5 questions, 75% users have

[1] In brief, SRCC assesses how well the relationship between two variables can be described using a monotonic function [5]

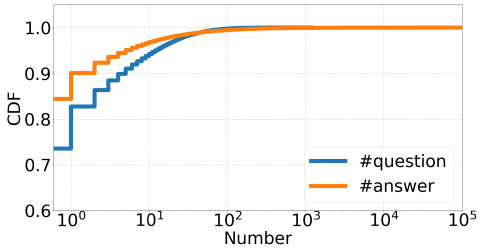

FIGURE 5. CDFs of the number of questions raised and answered by each user.

never raised any question. On the other hand, while 3% (9,583) users have answered more than 10 questions, 85% users have never answered any questions.

Meanwhile, we find that 71% (217,095) users have not raised nor answered any questions. We conjecture that they only join the community to learn from the questions raised by other users. On the other hand, 4.85% (14,803) users have only answered questions. They are possibly the supportive experts who join the community to help others.

b: DISCUSSION

For question routing, the highly skewed activity level of the users, as revealed by the above results, needs to be taken into account. For example, questions routed to active users are more likely to get timely answers. And routing questions to less active users, if performed accurately, will motivate them to provide an answer and hence increase the number of received answers.

3) CROPS

In this section, we examine the basic characteristic of the crops. In Farm-Doctor, users can specify the crops they are interested in. When raise questions, they can choose to associate the question with a tag specifying related crops. Farm-Doctor in total contains 438 crops that have been followed by users or have been tagged in questions.

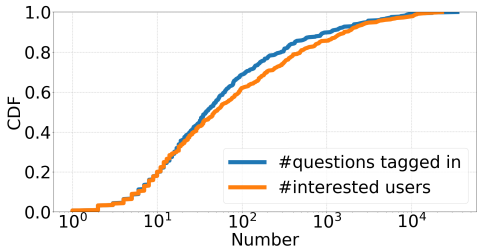

FIGURE 6. CDFs of the number of interested users and the number of questions a crop is tagged in.

Figure 6 shows the CDFs of the number of interested users and the number of questions a crop is tagged in, respectively. In total, 60% crops are interested to fewer than 100 users while the top 2% crops are interested to more than 10,000 users. On the other hand, 65% crops are tagged in fewer than 100 questions while the top 1% crops have more than 10,000 related questions. These results indicate an uneven distributed attention on the crops.

Discussion: While in Farm-Doctor, it is difficult to infer topics from the limited textual information, the crops followed by the users (for both users who raised the questions and the answerers) and tagged in questions provide rich side information for the question routing problem, which we will explore later in Section 5.

III. PRELIMINARY

In this paper, with n users $U = \{u_1, \ldots, u_n\}$ and m newly asked questions $Q = \{q_1, \ldots, q_m\}$, we predict whether the user $u_i, \forall i \in \{1, \ldots, n\}$ will answer the question $q_j, \forall j \in \{1, \ldots, m\}$. And then we can recommend the question $q_j, \forall j \in \{1, \ldots, m\}$ to the potential answerers. We perform recommendation tasks based on heterogeneous information network and network embedding, which can be define as following:

Definition 1 (Heterogeneous Information Network): Heterogeneous information networks contain multiple types of nodes or multiple types of edges, which can be defined as $G = (V, E)$. Through a node type mapping function $\phi : V \rightarrow T_v$ and an edge type mapping function $\psi : E \rightarrow T_e$ can obtain the type of each node and each edge, respectively. And must have the $|T_v| > 1$ or $|T_e| > 1$.

For example, as shown in Figure 1 (a), we model a Q&A community as a heterogeneous information network. It contains two node types (question and user) and three edge types (question-user answering, user-user answering, and question-question similarity).

Definition 2 (Network Embedding): Network embedding utilize algorithms to represent nodes in the network G with a low-dimensional dense vector space $X \in R^{|V|*d}, d \ll |V|$, which can maintain the relevant structure and characteristics of the original network.

In the matrix $\boldsymbol{X}$ (the output of the network embedding), each row corresponds to the representation of a node. As is shown in Figure 1 (b) and Figure 1 (c), The representations of all nodes have the same dimensionality, which can be used as features input into machine learning for classification, clustering and link prediction tasks.

IV. THE PROPOSED MODEL

In this section, to solve the question routing problem in Agricultural Q&A communities where the questions lack text descriptions, we propose a heterogeneous graph neural network model named HeSAC that consists of three parts: sampling, aggregation, concatenation.

A. LEARNING ALGORITHM

We summary our algorithm in Algorithm 1. In the input of the model, *G(V,E,Label)* means a heterogeneous information network with nodes set V, edges set E, and node type set *Label* (we have two node types: user (labeled u) and question (labeled q)), the parameter of K denotes the number of layers, $W^k, \forall k \in \{1, \ldots, K\}$ are the weight of the model,

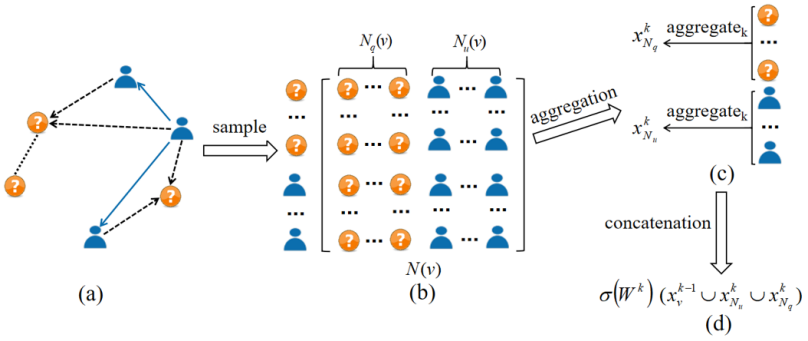

FIGURE 7. The overview of the HeSAC algorithm.

Algorithm 1 HeSAC Embedding Generation Algorithm

Input: Graph $G(V,E,Label)$, $Label \in u, q$; depth K;
weight matrices $W^k, \forall k \in 1,\ldots,K$; $non-linearity\ \sigma$;
aggregation function $AGGREGATE_k, \forall k \in 1,\ldots,K$;
Output: Vector representations $x_v, \forall v \in V$
$x_v^0 \leftarrow h_v, \forall v \in V$
for $k = 1 \ldots K$ do
  for v $\in V$ do
    $x_{N_u(v)}^k \leftarrow AGGREGATE_k((x)_u^{k-1}, \forall u \in N_u(v))$;
    $x_{N_q(v)}^k \leftarrow AGGREGATE_k((x)_q^{k-1}, \forall q \in N_q(v))$;
    $x_v^k \leftarrow \sigma(W^k.CONCAT(x_v^{k-1}, x_{N_u(v)}^k, x_{N_q(v)}^k))$
  end
  $x_v^k \leftarrow x_v^k/||x_v^k||_2, \forall v \in V$;
end

and $AGGREGATE_k$, $\forall k \in \{1,\ldots,K\}$ are the aggregation function. Next, we will introduce the model in three parts, sampling, aggregation, and concatenation.

1) SAMPLING

In our network, each node has a type label. As shown in Figure 7 (b), for each node, we will uniformly sample its different types of neighbors to obtain a fixed-size matrix $N(v)$. In this paper, there are two types of matrixes: the neighbors matrix N(v) contains neighbors of user type $N_u(v)$ and question type $N_q(v)$. In each iteration k, we will sample different number of neighbors.

2) AGGREGATION

In this paper, the aggregation function is the mean operator. In each layer, we aggregate user neighbors $N_u(v)$ and question neighbors $N_q(v)$, respectively. At the k layer, the aggregation operator ($AGGREGATE_k$) for user neighbors and question neighbors as following:

$$x_{N_u}^k = \alpha_u \sum_{\forall u \in N_u(v)} x_u^{k-1}, \quad \forall k \in \{1,\ldots,K\} \tag{1}$$

$$x_{N_q}^k = \alpha_q \sum_{\forall q \in N_q(v)} x_q^{k-1}, \quad \forall k \in \{1,\ldots,K\} \tag{2}$$

where $\alpha_u = \frac{1}{|N_u(v)|}$ and $\alpha_q = \frac{1}{|N_q(v)|}$ for all nodes in the mean-based aggregator.

3) CONCATENATION

Finally, the model concatenates the node representation in the previous layer x_v^{k-1} with the aggregated user neighborhood vector $x_{N_u}^k$ and question neighborhood vector $x_{N_q}^k$. And fed the concatenated vector into sigmoid function. The final representations x_v is the outputs at depth K. Specially, in order to reduce the computational cost, the the weight matrix W^k of each layer is shared. The operator as following:

$$x_v^k = \sigma(W^k \cdot (x_v^{k-1} \cup x_{N_u}^k \cup x_{N_q}^k)), \quad \forall k \in 1,\ldots,K \tag{3}$$

B. MODEL TRAINING

In this model, we need to learn the output representations x_u and the weight matrices W^k. To learn those parameters, we used the following objective function.

$$O(x) = -log\sigma(x_u^T x_v) - \sum_{n=1}^{N} \cdot E_{v^n \sim p(v)}[log\sigma(-x_u^T x_{v^n})] \tag{4}$$

where σ is the sigmoid function, p is the negative sampling distribution, and N is the number of negative samples. The basic ideas of the objective function is to maximize the inner product of the positive examples and minimize the inner product of the negative examples. Finally, To optimize the objective function, we adopt the stochastic gradient descent as the optimizer in our implementation.

V. EXPERIMENTS

Although questions in Farm-Doctor lack textual information, the previous analyses revealed rich side information that could be leveraged for the question routing problem. In this section, to solve the question routing problem in Farm-Doctor, we model Farm-Doctor as a heterogeneous information network (HIN) that incorporates different types of nodes and edges. Then, we utilized the HeSAC model to learn the low-dimensional vector representations of the nodes. Finally, we use the learned representation vectors as input features and build machine-learned classifiers to rank and to recommend the potential answerers.

A. EXPERIMENT SETUP

The HIN model we proposed for Farm-Doctor is shown in Figure 8, which consists of three types of nodes, i.e., users, questions, and crops, and three types of edges representing a variety of relationships including user-question answering, user-user answering, question-question similarity.

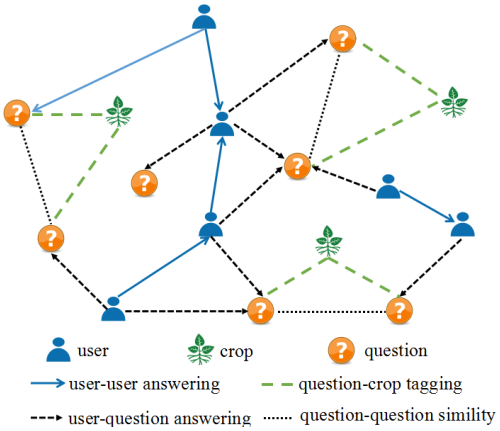

FIGURE 8. The heterogeneous information network model for Farm-Doctor.

For question routing, our task is to predict potential answerers for the newly posted questions. To this end, we divide the training dataset and the testing dataset according to time. The former contains the questions raised in the year of 2017 and the latter contains the questions raised in January and February 2018, which are the newly posted questions that need to be recommended to potential answerers.

TABLE 2. Statistics of the training dataset.

	Type	Count
44,387 nodes	Question	14,159
	User	30,228
2,904,969 edges	Question-user answering	206,958
	User-user answering	206,958
	Question-question similarity	2,677,353

We run HeSAC model on the network that consists of all the relationships in the training dataset and the question-question similarity relationships for the newly posted questions. Table 2 shows the statistics of training datasets. In total, the network contains[2] 44,387 nodes, representing 14,159 questions and 30,228 users, and 2,904,969 edges, representing 206,958 question-user answering relationships, 206,958 user-user answering relationships, and 2,677,353 question-question similarity relationships. We set the initial vector of each node to 128 dimensions, which is commonly used in neural network structures for similar tasks. After concatenating (as specified in Equation 3),

[2]Amazon-book: 70,679 users, 24,915 items, and 847,733 interactions, density: 0.0481%. Last-FM: 23,566 users, 48,123 items, and 3,034,796 interactions, density: 0.267%. Yalp2018: 45,919 users, 45,538 items, and 1,185,068 interactions, density: 0.0567%

TABLE 3. Vector functions.

Functions	Hadamard	average
Description	$\vec{v}_{1i} * \vec{v}_{2i}$	$\frac{\vec{v}_{1i}+\vec{v}_{2i}}{2}$

the final output feature vector of the HeSAC model, for each user and each question, is represented as 384 dimensions (given that our neural network contains three layers). We consider both the average and the Hadamard operator for generating the edge feature vectors according to the corresponding two nodes. The equations are shown in Table 3. Specifically, the Hadamard function is the element-wise multiplication of two vectors. The average function calculates the centroid of two vectors in the latent space. The Hadamard operator performs better. Hence all results reported here were obtained using the Hadamard operator.

Once we have obtained the representations of each node in the training dataset, we take the question routing as the downstream task which we model as a classification problem. More specifically, we use the learned representations as the input futures. And then we apply linear Support Vector Machine (SVM) to classify whether a question-answering relationship will be established between a user and a newly posted question. After the SVM model is trained, it is used to predict which users will answer questions occurred in the testing dataset.

More specifically, we label the user-question answering relationships established in January and February 2018 as the positive examples, and we use their neighbours within 2 hops as the negative examples. An example of the construction of the positive and the negative examples is depicted in Figure 9. In our experiment, to have enough information on the questions in analysis, we have only considered question with more than 5 replies from the testing dataset, which in total contains 680 questions.

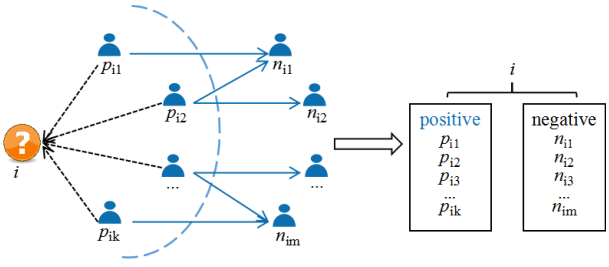

FIGURE 9. An example of positive and negative edge divide.

B. EVALUATION

Classic representation learning methods, especially for the recommendation problem, often rely on matrix factorization [6] that factorizes the relationship matrix into low rank latent matrices and the nodes into low-dimensional vectors. Here, we compare the performance of HeSAC models with the following two matrix factorization methods and one GNN model.

TABLE 4. Performance evaluation of question routing.

	precision@5	recall@5	precision@10	recall@10	precision@20	recall@20
PMF	0.4434	0.3513	0.4429	0.4442	0.4427	0.4475
CMF	0.4617	0.3801	0.4593	0.4515	0.4558	0.4742
GraphSAGE	0.4870	0.3892	0.4751	0.4758	0.4625	0.5113
HeSAC	**0.5191**	**0.4132**	**0.4873**	**0.5162**	**0.4695**	**0.5341**

- *PMF*: Probabilistic Matrix Factorization [7] is the basic matrix factorization method using only user-item matrix. In our experiment, the matrix is constructed in a way that, when a user i answers a question j, the corresponding element a_{ij} in the matrix is set to 1.
- *CMF*: Collective Matrix Factorization [8] is a matrix factorization model that jointly factorizes different types of relations in HIN and shares the latent factor of same node types in different relations. In our experiment, we utilize a user-question matrix (constructed in the same way as in the above PMF experiment), a user-user matrix (with an element a_{ij}=1 when a user i answers a question that asked by user j), and a question-question matrix (with an element a_{ij}=1 when a question i and a question j have similar crop).
- *GraphSAGE*: GraphSAGE [9] is a graph neural network model, and can generate the embedding of unseen node. But it designs for the homogeneous graph model.

We use the top-N precision (precision@N) and the top-N recall (recall@N) as metrics for the evaluation, which are defined as follows respectively:

- *Top-N precision* (precision@N): the percentage of top-N ranking results that hit the ground truth.

$$precision@N = \frac{\#hits\ in\ top\ N}{N}$$

- *Top-N recall* (recall@N): the fraction of the ground truth that are ranked in the top-N returned results.

$$recall@N = \frac{\#hits\ in\ top\ N}{\#positive\ examples\ in\ the\ ground\ truth}$$

C. RESULTS

The depth K is set to 2. In the first layer sampling, each node samples 10 user nodes and 5 question nodes from its neighbourhood. In the second layer sampling,each node samples 20 user nodes and 10 question nodes from its neighbourhood. To compare model performance, we set the node vector in all models to 384 dimensions. We have considered different values of N for our performance metrics, i.e., the top-N precision and the top-N recall. In addition, it should be noted that PMF only uses the user-question answering relationships and performs as the baseline for comparison. CMF uses user-question answering, user-user answering, and question-question similarity relationships. The GraphSAGE model uses user-question answering, user-user answering, question-question similarity relationships, but doesn't distinguish the type of the nodes. The HeSAC model uses all the relationships. The results are shown in Table 4. Taking the top-N precision and the top-N recall as the performance metrics, we have also evaluated the performance of our models for different values of N in Figures 10 and 11, respectively. From the results, we have four interesting findings as follows.

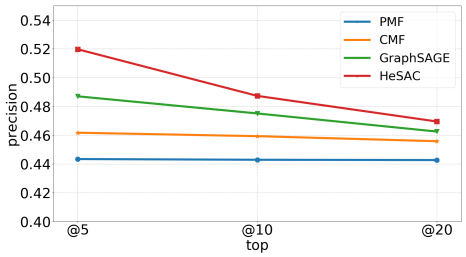

FIGURE 10. Top-*N* precision for difference models.

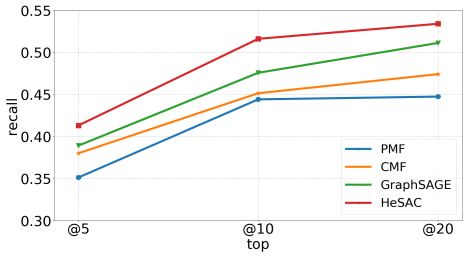

FIGURE 11. Top-*N* recall for difference models.

Firstly, CMF, GraphSAGE, and HeSAC, which contain heterogeneous information, significantly outperform PMF. This result indicates that the heterogeneous relationships we modeled have a significant influence in the question routing problem.

Secondly, we find that the GraphSAGE model and the HeSAC model perform better than CMF. This result suggests that the graph neural network model can get more accurate embedding for the nodes. In addition, both graph neural network models perform reasonably well, indicating that even without using any textual information, we can still accurately route the questions to the potential answerers.

Thirdly, we find that HeSAC performs constantly better than GraphSAGE, with 3% to 8% performance improvements. Nevertheless, as HeSAC is specially designed for HIN, one would expect it to achieve more notable improvements over graph neural network models designed for homogeneous networks (such as GraphSAGE).

Finally, for the HeSAC, as N increases from 5 to 10, the precision score decreases by 6%, but the recall score increases by 10%. AS N increases from 10 to 20, the performance of precision decreases by 4%, and the performance of recall only increases by 3%. Therefore, in order to achieve better recommendation results, the value of N should be set reasonably.

These results indicate that our work can provide insights into the question routing problem. Once the question is raised, we can use our model routing it to the users who are highly likely to answer it to ensure it can be answered in the short time.

VI. RELATED WORK

A. QUESTION ROUTING

Question routing recommends potential answers for each question, which is the focus of this work. This topic has been extensively studied before. Three comprehensive surveys can be found in [10]–[12]. Earlier works on question routing are mainly based on the past answering activities of the users. Li and King [3] utilized the query likelihood language model based on answerers' performance profile to estimate each answerer's expertise, and they further incorporate the category information [4]. Yang *et al.* [2] proposed a probabilistic Topic Expertise Model which used tagging information and voting information to learn topical expertise estimation. They also proposed CQArank model that combines user topical expertise estimation and user authority derived from link analysis to find experts with both similar topical preference and high topical expertise. Zhao *et al.* [13] considered the problem of expert finding from the viewpoint of missing value estimation. The missing value indicates the quality of users on answering the questions. Nobari *et al.* [14] proposed two models, the first model is based on a statistical approach and the second one is a word embedding model, to improve the matching between expert finding queries and answers. Based on both past question-answering activities and an user-to-user graph, Zhao *et al.* [15] proposed a Graph Regularized Latent Model (GRLM) to tackle the problem of cold-start expert finding.

The above methods all have utilized the rich textual information associated with the questions, which are not applicable to the question routing problem in Farm-Doctor where questions contain very limited texts. Closest to our work, a few studies have focused on routing questions without using textual information. Zhang *et al.* [16] proposed a graph model constructed by the question-answer relationship to compute the expertise scores of users via the PageRank algorithm [17]. Jurczyk and Agichtein [18] also proposed a graph model based on the user-user answering relationship, which used as input to the HITS algorithm [19] for predicting experts. For improving authority ranking, Zhu *et al.* [20] provided the extended category link graph which considered the category relevance of questions, and utilized a link analysis method which is based on Topical Random Surfer model for ranking user authority. Yang and Manandhar [21] factorized a user-tag expertise matrix through probabilistic matrix factorization [7] to obtain the latent feature vectors for the users and the tags. For a newly posted question, the predicted expertise scores can be computed by the feature vector of its tag and the feature vectors of the potential answerers. In contrast, we adopt NRL models to learn latent representations and we show that NRL models outperform matrix factorization methods. Zhao *et al.* [22] proposed a heterogeneous network model which is composed of users relationships extracted from other social website and user-questions answering relationships. They used a fully connected neural network and the LSTM model to learn the representations of the users and the questions. Compared to their work, our work consider more realistic scenarios by using additional tag information.

B. NETWORK EMBEDDING

Network embedding represents nodes in the network as a low-dimensional vector [23], [24], [25], which can be used to node classification [26], link prediction [27], and clustering [28]. In previous research, the latent representations for the nodes are generated by matrix factorization, such as non-negative matrix factorization [29], probabilistic matrix factorization [7], and collective matrix factorization [8]. However, the computational cost of matrix factorization is very expensive. And the advent of deep learning algorithms provides new ideas for learning node representations. Mikolov *et al.* [30], [31] proposed the word2vec algorithm, which can learn the distributed representation based on the ''context'' of words. Perozzi *et al.* [32] got the ''context'' of each node in a network through truncated random-walks. And then utilized word2vec algorithm to learn the node representations. Grover and Leskovec [33] proposed a biased random-walk method that are comprehensively considers breadth-first and depth-first sampling. Dong *et al.* [28] proposed a network representation learning algorithm for heterogeneous networks.

Recently, many researchers have used graph neural networks to learn node representations, which aims to apply the deep neural network into the graph-structured data. Based on graph neural network model, Hamilton *et al.* [9] proposed an inductive learning method for large-scale networks, which can generate the representations of unseen nodes. They learned a aggregation function that generates node embedding from local neighbors. But this model is designed for homogeneous graphs. For heterogeneous graph models, Wang *et al.* [26] combine graph neural network with attention mechanism and meta-path learning node representations, but they only focus on learning one type of node in the heterogeneous graph. Fan *et al.* [34] presented a graph neural network framework for social recommendations, which utilized user-user relationship and user-item relationships. This model only dealt with two relationships and relied on the rating information. Wang *et al.* [35] proposed a graph neural network model that can make full use of the utility of the knowledge graph, which contains more attributes information. The above work does not apply to our network; the heterogeneous information network in Farm-Doctor is very

sparse. In this paper, we proposed a novel model to learn the representations of the user and the question.

VII. CONCLUSION

In this article, we conduct an analysis of the question routing problem in a CQA platform named Farm-Doctor that is exclusive for agricultural knowledge. We address the major challenge for routing questions in Farm-Doctor, i.e., limited textual information, by proposing a heterogeneous information network model that captures a variety of relationships and proposing a novel GNN model to accurately predict and to recommend potential answerers. We also provide a publicly available dataset that contains the whole knowledge-base of Farm-Doctor, including detailed information on over 300 thousand users, over 690 thousand questions, and over 3 million answers. Promising directions for future work include designing GNN models for weighted and directed HINs, proposing methods for routing high quality answerers instead of any answerers, and comparing and combining GNN models with text-based approaches on more CQA platforms.

REFERENCES

[1] Z. Chen, C. Zhang, Z. Zhao, C. Yao, and D. Cai, "Question retrieval for community-based question answering via heterogeneous social influential network," *Neurocomputing*, vol. 285, pp. 117–124, Apr. 2018.
[2] 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 *Proc. 22nd ACM Int. Conf. Conf. Inf. Knowl. Manage. (CIKM)*, 2013, pp. 99–108.
[3] B. Li and I. King, "Routing questions to appropriate answerers in community question answering services," in *Proc. 19th ACM Int. Conf. Inf. Knowl. Manage. (CIKM)*, 2010, pp. 1585–1588.
[4] B. Li, I. King, and M. R. Lyu, "Question routing in community question answering: Putting category in its place," in *Proc. 20th ACM Int. Conf. Inf. Knowl. Manage. (CIKM)*, 2011, pp. 2041–2044.
[5] C. Spearman, "The proof and measurement of association between two things," *Amer. J. Psychol.*, vol. 100, nos. 3–4, p. 441, 1987.
[6] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, pp. 30–37, Aug. 2009.
[7] R. Salakhutdinov and A. Mnih, "Probabilistic matrix factorization," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2007, pp. 1257–1264.
[8] A. P. Singh and G. J. Gordon, "Relational learning via collective matrix factorization," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2008, pp. 650–658.
[9] W. L. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Proc. Adv. Neural Inf. Process. Syst.*, Long Beach, CA, USA, Dec. 2017, pp. 1024–1034.
[10] I. Srba and M. Bielikova, "A comprehensive survey and classification of approaches for community question answering," *ACM Trans. Web*, vol. 10, no. 3, p. 18, Aug. 2016.
[11] X. Wang, C. Huang, L. Yao, B. Benatallah, and M. Dong, "A survey on expert recommendation in community question answering," *J. Comput. Sci. Technol.*, vol. 33, no. 4, pp. 625–653, Jul. 2018.
[12] S. Yuan, Y. Zhang, J. Tang, W. Hall, and J. B. Cabotć, "Expert finding in community question answering: A review," *Artif. Intell. Rev.*, vol. 53, no. 2, pp. 843–874, Feb. 2020.
[13] Z. Zhao, L. Zhang, X. He, and W. Ng, "Expert finding for question answering via graph regularized matrix completion," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 4, pp. 993–1004, Apr. 2015.
[14] A. Dargahi Nobari, S. Sotudeh Gharebagh, and M. Neshati, "Skill translation models in expert finding," in *Proc. 40th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, 2017, pp. 1057–1060.
[15] Z. Zhao, F. Wei, M. Zhou, and W. Ng, "Cold-start expert finding in community question answering via graph regularization," *Lect. Notes Comput. Sci.*, vol. 9049, pp. 21–38, 2015.
[16] J. Zhang, M. S. Ackerman, and L. Adamic, "Expertise networks in online communities: Structure and algorithms," in *Proc. 16th Int. Conf. World Wide Web (WWW)*, 2007, pp. 221–230.
[17] A. Borodin, G. O. Roberts, J. S. Rosenthal, and P. Tsaparas, "Link analysis ranking: Algorithms, theory, and experiments," *ACM Trans. Internet Technol.*, vol. 5, no. 1, pp. 231–297, Feb. 2005.
[18] P. Jurczyk and E. Agichtein, "Discovering authorities in question answer communities by using link analysis," in *Proc. 16th ACM Conf. Conf. Inf. Knowl. Manage. (CIKM)*, 2007, pp. 919–922.
[19] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *J. ACM*, vol. 46, no. 5, pp. 604–632, Sep. 1999.
[20] H. Zhu, H. Cao, H. Xiong, E. Chen, and J. Tian, "Towards expert finding by leveraging relevant categories in authority ranking," in *Proc. 20th ACM Int. Conf. Inf. Knowl. Manage. (CIKM)*, 2011, pp. 2221–2224.
[21] B. Yang and S. Manandhar, "Tag-based expert recommendation in community question answering," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, Aug. 2014, pp. 960–963.
[22] Z. Zhao, Q. Yang, D. Cai, X. He, and Y. Zhuang, "Expert finding for community-based question answering via ranking metric network learning," in *Proc. Int. Joint Conf. Artif. Intell.*, 2016, pp. 3000–3006.
[23] P. Goyal and E. Ferrara, "Graph embedding techniques, applications, and performance: A survey," *Knowl.-Based Syst.*, vol. 151, pp. 78–94, Jul. 2018.
[24] J. Zhou, G. Cui, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun, "Graph neural networks: A review of methods and applications," 2018, *arXiv:1812.08434*. [Online]. Available: http://arxiv.org/abs/1812.08434
[25] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A comprehensive survey on graph neural networks," 2019, *arXiv:1901.00596*. [Online]. Available: http://arxiv.org/abs/1901.00596
[26] X. Wang, H. Ji, C. Shi, B. Wang, Y. Ye, P. Cui, and P. S. Yu, "Heterogeneous graph attention network," in *Proc. World Wide Web Conf. (WWW)*, 2019, pp. 2022–2032.
[27] T.-Y. Fu, W.-C. Lee, and Z. Lei, "HIN2 Vec: Explore meta-paths in heterogeneous information networks for representation learning," in *Proc. ACM Conf. Inf. Knowl. Manage. (CIKM)*, New York, NY, USA, 2017, pp. 1797–1806.
[28] Y. Dong, N. V. Chawla, and A. Swami, "Metapath2vec: Scalable representation learning for heterogeneous networks," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, New York, NY, USA, 2017, pp. 135–144.
[29] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, Oct. 1999.
[30] T. Mikolov, K. Chen, G. Corrado, and J. rey Dean, "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*. [Online]. Available: https://arxiv.org/abs/1301.3781
[31] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst.*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds. Red Hook, NY, USA: Curran Associates, 2013, pp. 3111–3119.
[32] B. Perozzi, R. Al-Rfou, and S. Skiena, "DeepWalk: Online learning of social representations," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, New York, NY, USA, 2014, pp. 701–710.
[33] A. Grover and J. Leskovec, "Node2vec: Scalable feature learning for networks," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, New York, NY, USA, 2016, pp. 855–864.
[34] W. Fan, Y. Ma, Q. Li, Y. He, E. Zhao, J. Tang, and D. Yin, "Graph neural networks for social recommendation," in *Proc. World Wide Web Conf. (WWW)*, New York, NY, USA, 2019, pp. 417–426.
[35] X. Wang, X. He, Y. Cao, M. Liu, and T.-S. Chua, "KGAT: Knowledge graph attention network for recommendation," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, New York, NY, USA, 2019, pp. 950–958.

XIAOXUE SHEN received the B.S. degree in the Internet of Things from Henan Normal University, China, in 2013. She is currently pursuing the master's degree with the Computer Science Department, China Agricultural University. Her research interests include complex network analysis and data mining.

ADELE LU JIA received the B.S. degree from the Harbin Institute of Technology, China, in 2007, the M.Phil. degree from The Chinese University of Hong Kong, in 2009, and the Ph.D. degree from the Delft University of Technology, The Netherlands, in 2013. She is currently an Associate Professor with the Computer Science Department, China Agricultural University. Her research interests include complex network analysis and data mining.

SIQI SHEN received the B.S. and M.S. degrees from the National University of Defense Technology, China, in 2007 and 2009, respectively, and the Ph.D. degree from the Delft University of Technology, The Netherlands, in 2015. He is currently an Assistant Professor with the National Laboratory for Parallel and Distributed Processing, National University of Defense Technology. His research interests include computer networks and data mining.

YONG DOU was born in 1966. He received the B.S., M.S., and Ph.D. degrees in computer science and technology from the National University of Defense Technology, in 1995. He is currently a Professor and a Ph.D. Supervisor and holds a Senior Membership at the China Computer Federation. His research interests include high-performance computer architecture, high-performance embedded microprocessor, reconfigurable computing, and bioinformatics.