Journal of Nonlinear and Convex Analysis
Volume 21, Number 7, 2020, 1487–1496

RESEARCH ON KNOWLEDGE QUESTION ANSWERING SYSTEM FOR AGRICULTURE DISEASE AND PESTS BASED ON KNOWLEDGE GRAPH

YINGCHUN XIA, NING SUN, HUI WANG, XIAOHUI YUAN, LICHUAN GU*, CHAO WANG, AND QIJUAN GAO

Abstract. A large increase of agricultural data results in the emergence of redundant information on the Internet. Users can not meet the results which are search by matching keywords or analyzing simple semantic. This paper proposes a knowledge question answering (QA) method to answer questions directly and succinctly in the filed of agriculture disease and pests. The methods is based on entity linking, external knowledge and similarity calculation. It prunes inappropriate entities and relations through entity linking, and uses external knowledge to improve recall rate. In addition, questions are represented by the sum of word vectors. Entities in knowledge graph are trained into vectors. The final answer is obtained by computing similarity between question and answer. A Chinese knowledge QA system is designed and implemented by using the proposed method. In the experiment, the dataset is generated according to RDF triple and seed questions. The experimental results show that the proposed method outperforms other state-of-the-art algorithms.

1. Introduction

At present, there are a wide range of big data sources in the agricultural field. Data representation, storage, organization and management are different. Information resources are highly dispersed and disorderly, which greatly affects the efficiency of users access to information and the degree of resource sharing [6, 9]. It is current research hotspot to search for more concise and accurate information. The QA system based on knowledge graph is a key technology to obtain accurate information for users questions [2, 17]. There have been some mature researches in many fields such as medical and financial. However, it is still starting stage in the agricultural field. Therefore, this paper starts with the knowledge graph of Chinese agricultural pests, and proposes a knowledge QA method based on entity link, external knowledge base and similarity calculation. The main contributions of this paper are as follows. 1) Entity linking is used to prune unrelated entities and relations.

2010 *Mathematics Subject Classification.* 97R40.

Key words and phrases. Agricultural pests, knowledge graph, knowledge question answering.

*Corresponding author.

This work is partially supported by the National Natural Science Foundation of China under Grant (31771679, 31671589), the Anhui Foundation for Science and Technology Major Project, China, under Grant (16030701092, 18030901034, 201904e01020006), the Key Laboratory of Agricultural Electronic Commerce, Ministry of Agriculture of China under Grant (AEC2018003, AEC2018006), the 2019 Anhui University collaborative innovation project (sn:GXXT-2019-013), the Hefei Major Research Project of Key Technology (J2018G14).

External knowledge base is used to improve Recall. 2) A knowledge QA system of agricultural pests is designed and implemented.

2. Related works

The QA system is a technique to answer natural language questions and provide users with concise and clear answers by locating, extracting and expressing corresponding answers in knowledge base [1, 5, 7, 8, 10, 11]. At present, the mainstream methods of natural language knowledge QA based on knowledge graph can be divided into four categories. 1) Template-based methods, such as TBSL [14]. TBSL answers a question by automatically translating questions into SPARQL templates and uses dictionary to map vocabulary in natural language into SPARQL components. These methods always have high accuracy and answer question quickly, but need to take a lot of time to build a large scale template base. 2) Methods based on graph exploration, such as Treo [4]. In additional, Shin et al. proposed Predicate Constraints based Question Answering (PCQA), which focused on relation query [13]. It answered questions by prunes inappropriate entity/relation matchings and generating query graph. 3) Methods based on semantic analysis. This category usually using learning model analyzers by using supervised learning method. L. Sang et al. proposed a generic Multi-modal Multi-view Semantic Embedding (MMSE) framework via a Bayesian model for question answering to deal with the multi-view property and sparse property of question answer pairs [12]. Semantic models usually rely on manual marking. The results are controllable, but has expensive cost. 4) Methods based on deep learning. Recently, Yang et al. proposed a novel Multi-task and Knowledge enhanced Multi-head Interactive Attention network for Community Question Answering, which studied an advanced deep neural network to represent questions and using external knowledge to help identify background information [16]. Embedding is flexible and can be used to deal with diverse questions. But neural network is unexplained and its results can not be controlled.

This paper combines graph and deep learning for question answering system. Word2Vec is used to analyze questions and graph is used to prune inappropriate entities and relations. Questions and Knowledge graph are represented by using embedding. The final answer is obtained by computing similarity.

3. Algorithm

3.1. **Question Analysis.** Knowledge QA usually has three steps, including questions analysis, candidate answers extraction and answers sequence. Analysis of natural language questions is an important step. The mainstream methods entirely involve semantic analysis techniques in question analysis. Relevant techniques have been developed more maturely, but implementation process is complex. Meanwhile, question analysis is not focus in this paper. Therefore, in order to comply simplification principle, this paper simplifies problem analysis. Rule matching strategy is adopted and Chinese question templates are stipulated. For example, *"What is the control method of XX?", "When does XX usually occur?"* and so on. Thereby, triples of entity and entity relationships can be directly extracted by designing specific questioning methods.

3.2. **Question Definition.** According to the questioning rules designed in Section 2.1, answering a question can be translated into a query for triplet. For example, there is a question *"What are symptoms of rice bakanae disease?"*. Question triplet *q(rice bakanae disease, symptoms, ?)* can be extracted from it. For such a problem triplet that subject and predicate have been given, answer can be obtained by using a graph query algorithm. In this paper, definition of knowledge QA can be given as follows. Given a question q, the answer a obtained from knowledge base is text. For Example.

q: What are symptoms of rice bakanae disease?

a: The diseased rice grains often do not germinate or can not be unearthed after sowing.

If same attribute of an entity have multiple values, or there are multiple entities with the same relationship, multiple $\langle Q, A\rangle$ pairs will be generated according to triplet. The QA based on knowledge graph will eventually return an exact answer, so these answers are needed to be sorted. Assume all the answers be denoted as set A and scoring function is $S(\bullet)$. The answer with the highest score can be expressed as formula (3.1).

$$\hat{a} = \underset{a\in A}{max} S(q, a) \tag{3.1}$$

Thus, the final answer is the answer that makes scoring function $S(q, a)$ be the largest. For multiple answers that may appear, this paper lists all answers according to their scores. Simultaneously, entity graph of core entity in question is given.

3.3. **Algorithm Architecture.** Figure 1 is flow chart of the knowledge QA algorithm in this paper. The specific steps are as follows.

Step1: Identifying core entity e_0. Core entity e_0, which is resource v_0 that may be mapped in the knowledge graph, is identified by using entity identification technology.

Step2: Reducing scope of query. The entity linking based on topic model and graph (ELTMG) algorithm is used to map core entity e_0 to the corresponding resource v_0 in local knowledge graph K_1. If there is no suitable resource, Chinese DBpedia K_2 is used as the supplementary knowledge base. The local knowledge graph is agricultural pests knowledge graph which constructed according to crop pest database in website of China National Agricultural Science Data Sharing Center.

Step3: Constructing an entity graph. The resource, which is mapped by core entity in knowledge graph, is used as the starting point. Then, entity graph is constructed by using breadth-first search to search other associated entities in knowledge graph.

Step4: Querying answer. All entities and entity relations has been represented in vectors through using TransE to train them. Vocabulary in question is represented in vectors by using Word2Vec. This paper query answer in entity graph

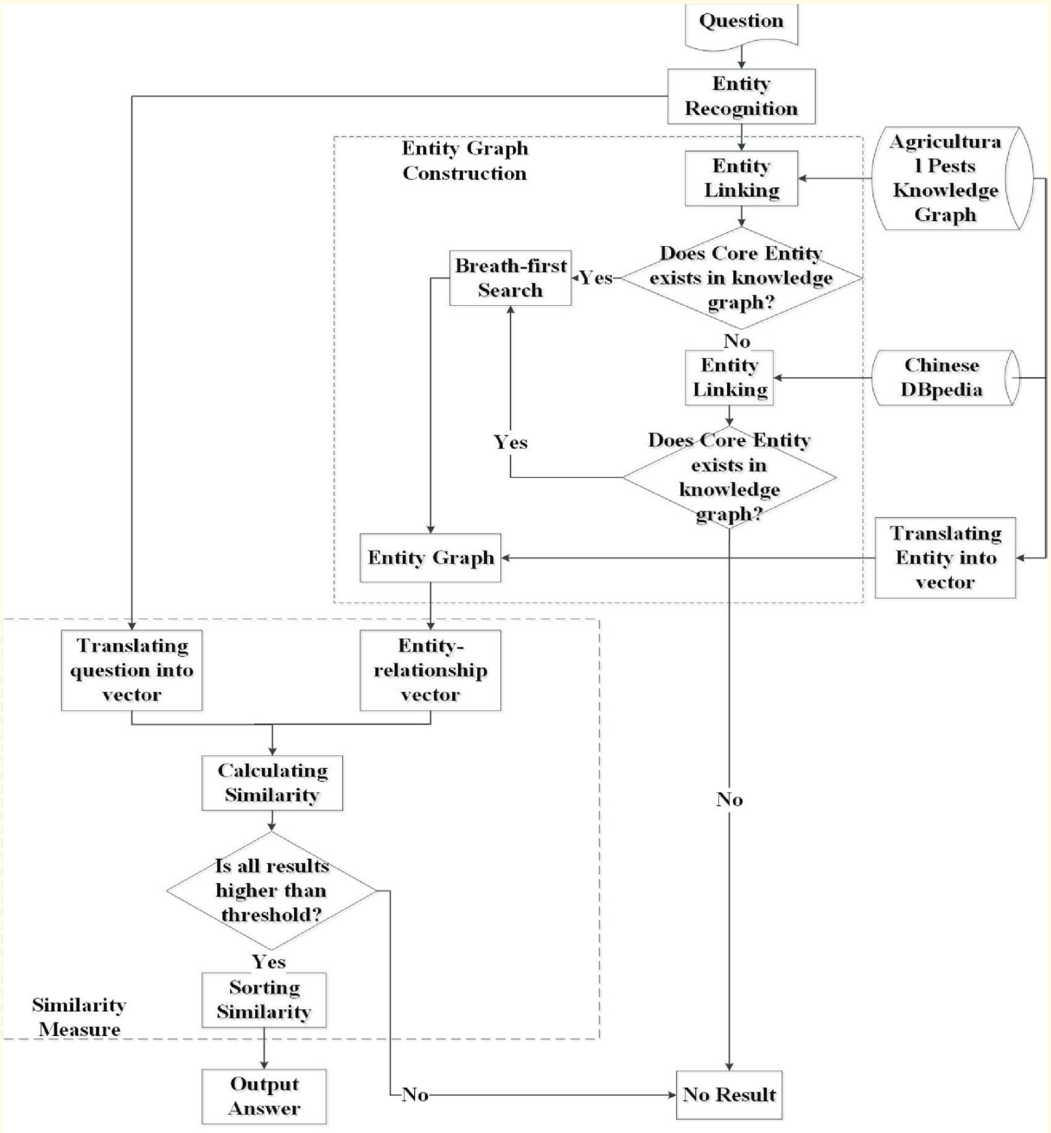

Question
Entity Recognition
Entity Graph Construction
Entity Linking
Agricultural Pests Knowledge Graph
Breath-first Search
Yes
Does Core Entity exists in knowledge graph?
No
Entity Linking
Chinese DBpedia
Yes
Does Core Entity exists in knowledge graph?
Entity Graph
Translating Entity into vector
Translating question into vector
Entity-relationship vector
Calculating Similarity
No
Is all results higher than threshold?
Yes
Sorting Similarity
Similarity Measure
Output Answer
No
No Result

Figure 1. Processing of knowledge QA.

by computing similarity between vocabulary vector and sum of entity relation pair vector.

3.4. **Entity Graph.** Construction of entity graph is designed to narrow scope of answer query. The process of entity graph construction can be described as follows.

Step 1: Constructing directed graph $G_{k_1} = (V, E)$ and $G_{k_2} = (V, E)$ of two knowledge graph. All of resources in knowledge graph are vertices V and all of relations in knowledge graph are edges E. If any vertices x, y in G_k satisfy the condition $x, y \in V$, $(x, y) \in E$, then (x, p, y) is an RDF triple in knowledge graph.

Step 2: Constructing initial graph $G_0 = (V_0, E_0)$. V_0 just contains resource v_0. E_0 is empty.

Step 3: Extending graph. In order to query each vertex and edge which relates with v_0, The breadth-first algorithm is used to extend G_0. According to RDF triple (x, p, y) in knowledge graph, other vertices and edges, which are related with v_0, can be searched and added into entity graph G_0. Defining the extended operation of graph $G_i = (V_i, E_i), i = 0, \ldots, d$ is $\rho(G_i) = G_{i+1} = (V_{i+1}, E_{i+1})$. Its extended rules are shown as formula (3.2) and formula (3.3).

$$V_{i+1} = V_i \sqcup y : \exists V_i \wedge (x, y) \in E \tag{3.2}$$

$$E_{i+1} = (x, y) \in E : x, y \in V_{i+1} \tag{3.3}$$

The final entity graph is obtained by performing operation with d times on initial graph G0.

3.5. **Similarity Measure.** This paper matches questions and answers by calculating similarity between word vector of question and entity-relationship vector in knowledge graph. Word vectors $\phi(q)$ of question is generated by using Word2vec. The vector representation of question $f(\bullet)$ is shown as formula (3.4).

$$f(q) = \sum \phi(q) \tag{3.4}$$

The entities and entity relationships in knowledge graph are trained using TransE to obtain vector representation. Corresponding to the entity graph G_d in section 3.4, assume $g(v_0, e_i)$ represents other entity vectors in entity graph G_d expect for v_0. TransE and word2vec are based on vector addition. The similarity cosine measure, which computes angle of vectors is more appropriate. Therefore, Similarity between entity vector and question vector is obtained by calculating the cosine value of two vectors. Scoring function is expressed as formula (3.5).

$$S(f, g) = cos(f(q), g(v_0, e_i)) \tag{3.5}$$

Answers are sort according to similarities. The final answer can be expressed as formula (3.6).

$$\hat{a} = max(cos(f(q), g(v_0, e_i))) \tag{3.6}$$

4. Experiments

4.1. **Datasets.** Knowledge QA lacks high-quality public datasets in Chinese, especially in agriculture filed. This paper generates datasets and artificially set seed questions based on RDF triples in agricultural pests knowledge graph to evaluate the proposed algorithms. According to the rules in section 3.1, RDF triples are processed and mapped into natural language questions. Some seed questions are shown in Table 1. For example, the third seed question in Table 1 can be mapped into the question *"When is the occurrence time of rice seedling disease?"*.

Table 1. Templates of seed problems.

Seed problems
What are symptoms of xxx?
What are controls method of xxx?
When does xxx usually occur?

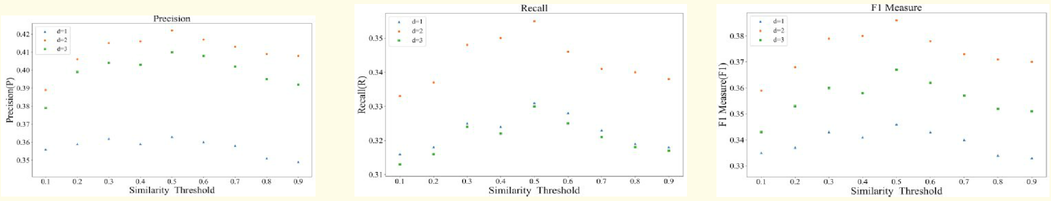
Dataset, which is mapped by knowledge graph, is shown in Table 2. Both the training set and the test set contain the questions and corresponding answers. Due to the limited number of entities in the local knowledge graph, size of the dataset is small, but it is sufficient to verify the effect. In the future, more Internet information will continue to be integrated.

Table 2. Datasets in disease and pest of agriculture crop.

Datasets	Training set	Test set
Agriculture pests data	1928	254

The parameters in this paper are mainly breadth search depth d and similarity threshold. d is selected from 1, 2, 3. Range of δ is set from 0 to 1 with a step size of 0.1.

4.2. **Results.** This paper evaluates the proposed method by using precision P, recall rate R and $F_1 - Measure$ F_1. According to parameter setting in this paper, training set is used to train the algorithm and obtain the optimal parameters. Test set is used to verify performance. In this experiment, if there are multiple correct answers to a question, the top one with the highest score is chosen as the correct answer.



(a) Precision (b) Recall (c) $F_1 - Measure$

Figure 2. Results in different parameters.

As shown in Figure 2, the value of precision, recall and $F_1 - Measure$ changes with the increasement of similarity threshold and search depth. The optimal parameters is $d = 2, \delta = 0.5$. When the optimal parameters, $F_1 - Measure = 0.386$. When search depth d is too small, number of entities in entity graph are limited, which may result in an failure search. With d increases, the probability of matching a correct answer increases. However, similar answers caused some interference in the answer query. Therefore, $d = 2$ is more appropriate. For similarity threshold, it is the opposite of search depth. When similarity threshold is too small, too many candidates affect efficiency and precision. And too large similarity threshold leads to loss of correct candidates. As shown in the results, it has the best value when similarity threshold equals to 0.5.

TABLE 3. Comparison in external knowledge.

With external base			With out external base		
P	R	F_1	P	R	F_1
0.431	0.357	0.391	0.342	0.296	0.317

Table 3 is comparison in external knowledge. The test experiment tests the proposed method in two situations. 1) knowledge base is only the agricultural pest knowledge graph. 2) knowledge base contains Chinese DBpedia as supplement. Comparing results of two situations, $F_1 - Measure$ of the second situation has increased by 7.4% than the first.

TABLE 4. Comparison with other algorithms.

Algorithm	$F_1 - Measure$
The proposed method	**0.386**
DEANNA	0.329
Aqqu	0.362

Table 4 Shows comparison with other algorithm. We reproduced DEANNA [15] and Aqqu [3] in our dataset using the code published by authors. The proposed method has the maximum avarage $F_1 - Measure = 0.386$. Bad performance of The other algorithms may be caused by the dataset which is in the filed of agricultural pests. And the proposed method can not have a good performance, because the agricultural knowledge graph has not enough knowledge and DBpedia is not a special knowledge graph. In future, we will study deeply to address these problems.

4.3. System.

4.4. **System Architecture.** The system mainly provides QA function in agricultural pests based on the above method. Figure 3 shows architecture of knowledge QA system in agricultural pests. It is mainly divided into four modules: front-end display module, background processing module, knowledge base construction, and QA module.

1) Front-end display module

The front-end display module visually presents functions to users. The query results and other results related with answer are displayed in this module.

2) Background processing module

This module is divided into two pieces, including question analysis and core entity recognition. The result is input of the QA module.

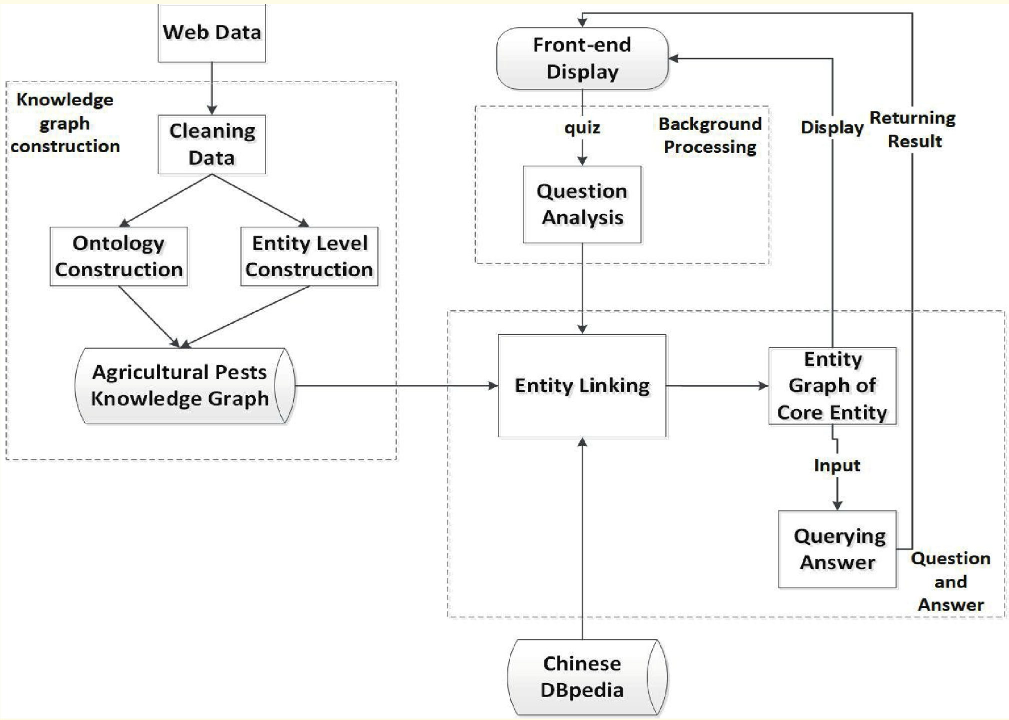


Figure 3. Architecture of knowledge QA in agriculture pests.

3) Knowledge Base Building Module

The knowledge base used in this paper is divided into two parts. One is agricultural pests knowledge graph which is constructed by using web crawler technology and D2R tools. Its data source is crop pest database under website of National Agricultural Science Data Sharing Center of China. To achieve visualization, a portion of the knowledge graph is stored in Neo4j graph database. The other is Chinese DBpedia.

4) QA module

The result of background processing module is take as input of this module. According to the proposed method, question input is processed and answer is queried. If answer exists, result is output in front-end display module. Otherwise, the front-end display module output *“No Result”*.

4.5. **Display.** This paper implements a simple Chinese QA system. Figure 4 shows the page for submitting questions. It contains question input box, answer display area and display area for core entity graph. When user inputs a question, such as *“What are the rice diseases?”*, answer and entity graph are returned and displayed on page. The specific result is shown in Figure 5. Answers are sorted by their scores. In experiments, the answer with maximum score is regarded as the right one. Answer candidates are shown in entity graph and listed in answer display area.

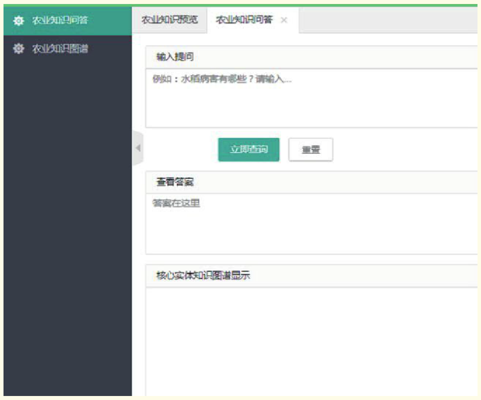
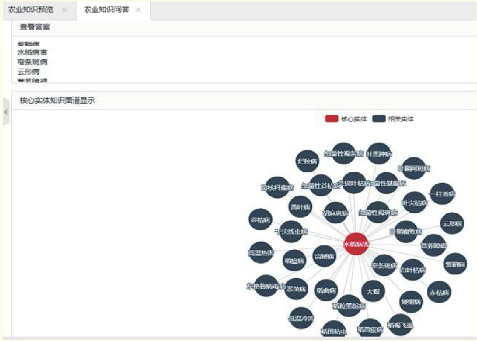


FIGURE 5. The index page of agricultural knowledge QA system.



水稻病害
窄条斑病
云形病

FIGURE 6. Page of answers.

5. CONCLUSIONS

In order to avoid useless answers obtained by traditional search methods, this paper proposes a knowledge QA method based on entity linking and external knowledge and similarity calculation to answer questions directly and precision. The entity link algorithm is used to reduce scope of query. To address the shortcomings of local agricultural pest knowledge graph, Chinese DBpedia is used as a supplement. Experimental results show that $F_1 - Measure$ of the proposed method is 39.1% and outperforms other state-of-the-art algorithms. Based on the proposed method, this paper designs and implements a knowledge QA system in agricultural pest and initially achieves agricultural knowledge service.

REFERENCES

[1] A. B. Abacha and P. Zweigenbaum, *MEANS: A medical question-answering system combining NLP techniques and semantic Web technologies*, Information Processing and Management **51** (2015), 570–594.

[2] A. Abdi, N. Idris and Z. Ahmad, *QAPD: an ontology-based question answering system in the physics domain*, Soft Computing **22** (2016), 1–18.

[3] H. Bast and E. Haussmann, *More accurate question answering on freebase*, in: Proceedings of the 24th ACM International Conference on Information and Knowledge Management, 2015, pp. 1431–1440.

[4] E. Curry, *Treo: best-effort natural language queries over linked data*, in: International Conference on Natural Language Processing and Information Systems, Springer-Verlag, 2011, pp. 286–289.

[5] B. Federico and P. Viviana, *Ontology-based affective models to organize artworks in the social semantic web*, Information Processing and Management **52** (2016), 139–162.

[6] L. L. Gu, *Research on Information Resources Integration and Service in Professional Field of Agriculture*, Beijing, 2016.

[7] Z. D. Han, *Question answering with a conceptual framework for knowledge-based system development "Node of Knowledge"*, Expert Systems with Applications **42** (2015), 5264–5286.

[8] A. J. Kumar, C. Schmidt and J. Köhler, *A knowledge graph based speech interface for question answering systems*, Speech Communication **92** (2017), 1–12.

[9] X. Liu, H. Y. Zheng, N. Q. Shi, Y. M. Liu and Y. Z. Lin, *Artificial intelligence in agricultural applications*, Fujian Journal of Agricultural Sciences **28** (2013), 609–614.

[10] W. Lu, J. Cheng and Q. Yang, *Question answering system based on web*, in: Fifth International Conference on Intelligent Computation Technology and Automation, IEEE, 2012, pp. 573–576.

[11] J. Peral, A. Ferrández, E. D. Gregorio, J. Trujillo,A. Maté and L. José, *Enrichment of the phenotypic and genotypic data warehouse analysis using question answering systems to facilitate the decision making process in cereal breeding programs*, Ecological Informatics **26** (2015), 203–216.

[12] L. Sang, M. Xu, S. Qian and X. Wu, *Multi-modal multi-view Bayesian semantic embedding for community question answering*, Neurocomputing **334** (2019), 44–58.

[13] S. Shin, X. Jin, J. Jung and K. Lee, *Predicate constraints based question answering over knowledge graph*, Information Processing and Management **56** (2019), 445–462.

[14] C. Unger, J. Lehmann, A. C. N. Ngomo and P. Cimiano, *Template-based question answering over RDF data*, in: International Conference on World Wide Web, 2012, pp. 639–648.

[15] M. Yahya, K. Berberich, S. Elbassuoni, M. Ramanath, V. Tresp and G. Weikum, *Natural language questions for the web of data*, in: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning,2012, pp. 379–390.

[16] M. Yang, W. Tu, Q. Qu, W. Zhou, Q. Liu and J. Zhu, *Advanced community question answering by leveraging external knowledge and multi-task learning*, Knowledge-Based Systems **171** (2019), 106–119.

[17] T. Yu, J. Liu and L. R. Jia, *Research on the construction of big knowledge graph for traditional Chinese medicine*, China Digital Medicine **10** (2015), 80–82.

Manuscript received March 25, 2019
revised June 12, 2019

Y. Xia
School of Information and Computer, Anhui Agricultural University, Hefei, 230036, China
E-mail address: 1415220245@qq.com

N. Sun
School of Information and Computer, Anhui Agricultural University, Hefei, 230036, China
E-mail address: zodomain@163.com

H. Wang
School of Information and Computer, Anhui Agricultural University, Hefei, 230036, China
E-mail address: 1835435117@qq.com

X. Yuan
Department of Computer Science and Engineering, University of North Texas, TX, 76203, USA
E-mail address: xiaohui.yuan@unt.edu

L. Gu
School of Information and Computer, Anhui Agricultural University, Hefei, 230036, China
E-mail address: glc@ahau.edu.cn

C. Wang
School of Information and Computer, Anhui Agricultural University, Hefei, 230036, China
E-mail address: warton_wang@qq.com

Q. Gao
School of Information and Computer, Anhui Agricultural University, Hefei, 230036, China
E-mail address: 181510179@qq.com