Agriculture Knowledge Graph Construction and Application

Hongchen Qin\textsuperscript{1, *} and Yiheng Yao\textsuperscript{2}
\textsuperscript{1}University of Chinese Academy of Sciences
\textsuperscript{2}Institute of Automation. Chinese Academy of Sciences
*Email: qinhongchen2019@ia.ac.cn

Abstract For the purpose of establishing vertical knowledge graph and auxiliary applications in the agricultural field, a set of agricultural knowledge graph construction methods, calculation frameworks and practical application systems are proposed. Firstly, the existing storage form and knowledge representation of knowledge in the agricultural field are integrated and regularized. On the basis of this data processing, the intelligent construction method of automatic and manual dual mode of knowledge graph in the agricultural field is proposed, and the key technology of entity relationship joint model to extract entity relationship and intelligent retrieval of irregular data. Then, similarity calculation will be used to perform entity knowledge fusion on knowledge graph in the agricultural field, making the graph more standardized, accurate and complete. A good graph is visualized and applied to the mainstream functions of intelligent question answering, which makes the whole system sort out the messy agricultural knowledge and apply it better to better assist learning and research.

Keywords: knowledge graph, Knowledge representation, intelligent search, intelligent construction, joint model, knowledge fusion, intelligent question and answer

With the continuous improvement of the efficiency and accuracy of the problem-solving requirements, traditional search engines and query methods have been unable to meet the efficiency and accuracy of knowledge acquisition in the agricultural field. Google proposed the "knowledge graph" technology in 2012\textsuperscript{[1]}. This technology stores structured knowledge, knowledge expansion, and efficiency and accuracy of knowledge query. There is a huge improvement over the more traditional search algorithms. At the same time, it understands the user's intention from the semantic level, and provides accurate answers after intelligent analysis, quickly and efficiently meets the needs of users, and provides research reference for the realization of intelligent knowledge services.

China is a large country with agriculture as the main industry. Therefore, building a knowledge graph in the agricultural field is an important driving force for China to improve agricultural productivity, promote research in the agricultural field, and analyze target relationships. As a technical system, knowledge graph is the representative progress of knowledge engineering in the era of big data. At the same time, knowledge graph is also the technical basis of user intention understanding, relationship interpretation, and user portrait.

Therefore, it is very important to construct agricultural knowledge graph.

1. Necessity of constructing agricultural knowledge graph

With the continuous development of Internet technology, massive agricultural data and information are surging like tide, which contains a lot of valuable knowledge. People hope to find an effective way to
discover, manage and use these massive amounts of agricultural knowledge, so we propose a complete system of knowledge graph in the field of agriculture. As a large-scale semantic network, it is the knowledge representation in the era of big data. One of the important ways is as a technical system, which is the representative progress of knowledge engineering in the era of big data. So it is very important to construct agricultural knowledge graph.

1.1 Integrating knowledge graphs to transform scattered data into structured knowledge
With the explosive growth of agricultural data, how to standardize the expression and management of agricultural data information resources is the prerequisite for the intelligent and three-dimensional development of knowledge service systems. This knowledge is complementary and mutually corroborating, so that the knowledge service feedback is more comprehensive and accurate, but these large amounts of data have large format differences, so it is urgent to carry out a unified structured expression study of massive unstructured data information, for the foundation of agricultural knowledge graph construction and efficient knowledge acquisition is laid. The knowledge graph makes these unstructured data finally stored in the form of "entity-relationship-entity" triplet, so that all unstructured and semi-structured data is converted into structured data, making the data both accurate and not redundant. Express the relationship between different entities.

1.2 Knowledge graph can fuse redundant data
Because the massive agricultural data obtained from the network is scattered and messy, after the unstructured and semi-structured data is converted into structured data through the model, it is inevitable that there will be different expressions of the same entity. There are a large number of multi-source heterogeneous data. The noise and redundant information need to be unified under the unified framework specification through knowledge fusion, through entity alignment, knowledge processing, and knowledge update, and then stored in the knowledge graph, which also includes entity linking, entity disambiguation and other technologies. The knowledge graph can integrate the same entity of different expressions into the same expression through an algorithm based on semantic similarity, thereby reducing the problem of data redundancy and improving search efficiency and accuracy.

1.3 Knowledge graph can mine knowledge and complete
Although the agricultural knowledge graph generated through the Internet of agricultural data has a huge amount of data, it is inevitable that the basic relationship will be missing for various reasons during the construction process, resulting in the construction of the agricultural knowledge graph is not so complete. Whether the knowledge graph is complete directly affects the use of this search engine, intelligent question answering system and other applications [2]. Because the data volume is too large, the manual completion method, that is, inefficient cost is not easy to control. Therefore, some scholars have developed some efficient algorithms to enable knowledge graphs to automatically complete knowledge mining. Solve the problem of lack of knowledge and other applications that affect search, question and answer.

1.4 The knowledge graph can infer existing knowledge at a logical level
Compared with traditional search engines, knowledge graph has a significant improvement in knowledge search reasoning[3]. The traditional search algorithm only makes the search through first-order logic and keywords in the text to display not completely accurate answers, while knowledge graph It is through the analysis of the multi-layer connected relationship between entities to infer the best matching answer. For example, if we search for "Who is Yao Ming's wife" through a traditional search engine, the traditional search engine can get an accurate answer, but if we search for "Who is Yao Ming's wife's father", the traditional search engine search results are "Yao Ming's wife", "Yao Ming's father", "wife's father" and other wrong answers, but if the reasoning search engine through knowledge graph, you can get the correct answer "Yao Ming's wife's father". Therefore, the knowledge graph can accurately infer when encountering problems of multi-layer logic to obtain accurate search answers[4].
2. Overall process of construction and application of agricultural knowledge graph

This paper proposes a complete set of agricultural knowledge graph construction system from scratch to specific application. It mainly includes four modules: knowledge joint extraction model, open source structured data extraction triplet, knowledge fusion based on word vector similarity calculation, and knowledge graph intelligent question and answer application. The overall flow of the system is shown in Figure 1.

As shown in Figure 1, first of all, collect massive data about agriculture through webpages, journals, papers and conferences, and store and express these data in the form of text. The data at this time is unstructured data.

Secondly, it is necessary to study the structured expression method of massive scientific and technological literature information based on deep neural networks; the core of the knowledge graph is the triplet, that is, two entities and their relationship. This is the cornerstone of the knowledge graph, and all subsequent processing including various machine learning, inference, mining, and intelligent retrieval algorithms are based on triples. Therefore, it is necessary to have a better mathematical description of the entity, relationship and triplet itself, that is, a more effective vector representation in space. According to the effective vector representation, we here use the created entity relationship joint model based on Bert + BiGRU + CRF to extract triplets.

Thirdly, combine a variety of machine learning theories and methods to construct a knowledge graph and a graph update mechanism; the knowledge graph in this project is different from the general knowledge graph. The entity relationship of this knowledge graph is complex, involving many factors and presenting many features of labeling. Therefore, it is proposed to use machine learning to accurately fuse entities. Whether it is the self-learning of knowledge graph or the reasoning of tacit knowledge, it involves a variety of machine learning theories and methods such as unsupervised analysis of incremental data, incremental learning, graph filling and so on.

Finally, research efficient knowledge graph storage and retrieval algorithms and easy interactive visualization methods to form a knowledge graph inference module that supports intelligent knowledge acquisition. The knowledge graph essentially connects the agricultural information entity data through a network of knowledge points. Users can make intelligent question and answer queries in natural language, graph user query statements to knowledge graph nodes, and perform certain analysis and reasoning on the subnets around the nodes to obtain the service knowledge that users need.

3 Knowledge joint extraction model
At present, as artificial intelligence research enters a stage of rapid development, in natural language processing, named entity recognition (NER) technology has been quite mature, the accuracy rate of entity extraction can reach 95% and above, but the extraction of relationships between entities does not have too good performance, most of the current relationship extraction technology is to build a relationship database, and then identify the entity to further determine which type of relationship in the relationship database belongs to this entity pair. This extraction method causes the relationship to be judged by the entity rather than the semantics of the sentence itself, so that the relationship of entity pairs cannot be obtained from the sentence. That is, the type of entity relationship depends more on the entity itself than the sentence. This paper adopts Bret + BiGRU + CRF as the main neural network and related algorithm model to jointly extract the entity and relationship of text information. The improved model in this paper is based on the improvement of *Joint entity recognition and relation extraction as a multi-head selection problem* [5], which makes the model extract the relationship according to the sentence while extracting the entity according to the sentence. The main flow of the model is shown in Figure 2.

![Figure 2 The main structure and process of the model](image)

### 3.1 Essentials of data preprocessing

This model requires the text to mark the subscript (id) of the word in the text, the label tag adopts the BIO strategy (bio), the entity relationship (no relationship is N) (relations), and the corresponding relationship subscript position (no relationship is Current subscripts) (heads), such as "The origin of rice is China", the label is shown in Figure 3.

---

| 0 | The | O | N | 0 |
|---|-----|---|---|---|
| 1 | origin | O | N | 1 |
| 2 | of | O | N | 2 |
| 3 | rice | B-COR | Source | 5 |
| 4 | is | O | N | 4 |
| 5 | China | B-LOC | N | 5 |

**Figure 3 Data annotation example**

After labeling the data, we need to preprocess the data. First, we need to read all the data and obtain the complete set of bios_set and relationship_set of the entity tags of all the data. Secondly, traverse the training data, change words to id, each sentence has the same dimension, and the insufficient ones are filled with 0, and finally encapsulated with scoringMatixHeads into the sentence.

The calculation method of scoringMatixHeads is:

First get the ids of the relations, and the id corresponds to the inside of the relations_set. For example, ‘rice’ corresponds to ‘[source]’, and source has a subscript of 3 in the relations_set, then the ids corresponding to the relations is 3.

Then iterate through the list of relations of the word correspondence relationship, and heads * len (relations_set) + relations. For example, ‘rice’ corresponds to relations 3, corresponding heads to [5],

---
and relations_set length to 10, then scoringMatrixHeads = [5 × 10 + 3] = [53].

3.2 The main process of named entity recognition
As shown in Figure 2, this article uses the currently popular Bert word embedding model to make the resulting sentence features more accurate. In named entity recognition, we use the BIO encoding scheme. In each sentence, we provide each word BIO coded to determine the entities in the sentence. Then the transferred word vector was passed to BiGRU. In this experiment, we chose the BiGRU neural network instead of BiLSTM because the training speed of BiGRU is faster and the performance of BiGRU is better than BiLSTM. After passing BiGRU, we connect a CRF layer to calculate the score in each word or corresponding BIO tag. The formula for calculating the score is:

\[ s^{(e)}(h_i) = V^{(e)} \text{relu}(U^{(e)} h_i + b^{(e)}) \] (1)

The e in the formula is used to mark the NER task, \( V^{(e)} \in \mathbb{R}^{(p \times l)} \), \( U^{(e)} \in \mathbb{R}^{(l \times 2d)} \), \( b^{(e)} \in \mathbb{R}^l \). Where d is the size of the BERT hidden layer, p is the total number of entity identification label types, and l is the width of BiGRU. Then define the score through CRF, the formula is:

\[ S(y_1^{(e)}, \ldots, y_n^{(e)}) = \sum_{i=0}^{n} s_{y_i^{(e)}} + \sum_{i=1}^{n-1} T_{y_i^{(e)}, y_{i+1}^{(e)}} \] (2)

Where \( S \in \mathbb{R} \), \( s_{y_i^{(e)}} \) is the predicted label score of the word, y refers to the predicted vector of the entity label, and T is a square transposed matrix, specifically Refers to the probability of conversion from one entity tag type to another entity comparison type.

In the end, each word of a given sentence or word's prediction overview at all given tags is defined as:

\[ \text{Pr}(y_1^{(e)}, \ldots, y_n^{(e)} | \theta) = \frac{e^{s(y_1^{(e)}, \ldots, y_n^{(e)})}}{\sum_{y_1^{(e)}, \ldots, y_n^{(e)}} e^{s(y_1^{(e)}, \ldots, y_n^{(e)})}} \] (3)

Finally, the Viterbi algorithm obtains the prediction mark with the highest score. Here we train the CRF layer by minimizing cross entropy.

3.3 The main process of relationship extraction
Through the previous neural network layer, we have obtained the entity label prediction of the word or word. We will embed the predicted label through the Embedding layer and splice it with the previously converted feature vector of the Bert layer. Next, we use the same relationship extraction formula as Joint entity recognition and relation extraction as a multi-head selection problem [5] to obtain the possible corresponding relationship and head vector of each word. The score between them is recorded as relation_scores, the formula is :

\[ s^{(r)}(z_i, z_o, r_k) = V^{(r)} \text{relu}(U^{(r)} z_i + W^{(r)} z_o + b^{(r)}) \] (4)

r in the formula is used to mark the relation extraction task, \( V^{(r)} \in \mathbb{R}^{(l)} \), \( U^{(r)} \in \mathbb{R}^{(2d+2b)} \), \( W^{(r)} \in \mathbb{R}^{(2d+b)} \), \( b^{(r)} \in \mathbb{R}^l \). Where d is the size of the BERT hidden layer, b is the size of the tag embedding, and l is the width of the BiGRU. The obtained scores are processed by the sigmoid layer function to obtain a set of "entity-relationship-entity" between different word vectors. In the training process, we use the Adam optimization function to optimize the loss.

4. Open source structured data extraction triplet
With the development of the network and the explosive increase of agricultural data, in order to build a deep agricultural knowledge graph, we not only automatically identify agricultural entities and relationships in unstructured data through artificial intelligence, we also use existing structured and Agricultural knowledge participates in the construction of the agricultural knowledge graph, making the agricultural knowledge graph more complete for better application in other areas. We mainly conduct
data mining from existing open knowledge graphs and official websites in the agricultural field. In the end, more than 120,000 entity nodes and more than 270,000 relationships were stored in these two ways.

4.1 Extract structured data based on existing open-source knowledge graphs in the open field

Based on the currently more mainstream open source knowledge graphs such as Knowledge Graph, DBpedia, Sogou "knowledge cube" and Baidu "knowledge", etc., through the related keywords, the relevant information in the graph is extracted, and after standardization processing, the accurate information is filtered. The agricultural knowledge is stored in the RDF standard of "entity-relationship-entity".

4.2 Crawling massive agricultural data through web pages

By using crawler technology to crawl data in the agricultural field in a reasonable and legal manner, by parsing the network address and analyzing the structure of the web page, the crawled data is normalized, and the knowledge is finally "entity-relationship-entity" RDF. The standard is the template for text storage.

5 Knowledge fusion based on word vector similarity calculation

The unstructured data is converted into RDF-standard triplet structured data through the entity relationship joint model and the RDF-standard triplet structured data is extracted from the open source structured data to contain entity relationships and entity attributes. The goal of the information, however, there is a lot of redundancy and error information in the information. The relationship between data is also flat and lacks hierarchy and logic, so it is necessary to clean and integrate it [6]. Knowledge fusion is mainly for entity linking and knowledge merging, merging the same entity in multi-source data, so that different types of attributes of different source data of the same entity are placed on the same entity.

In this paper, the TF-IDF algorithm based on word vectors and semantics and the cosine distance are used to calculate the similarity for knowledge fusion. First, we first obtain the feature vectors of different entities through the TF-IDF algorithm. Second, we calculate the cosine similarity between different entities through the cosine distance. When the similarity is greater than the threshold we set at the beginning, we think that the two are the same entity.

5.1 Principle of TF-IDF algorithm

TF-IDF is the Term Frequency-Inverse Document Frequency algorithm. TF-IDF is a statistical method that obtains the feature vector of each word through the word frequency and the reverse word frequency of each word in the text.

TF represents the frequency of keywords in the text. This word number is usually normalized to prevent it from biasing to long text. The formula is:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

(5)

Where \(n_{i,j}\) is the number of occurrences of this word in this text, and the denominator is the sum of the number of occurrences of all words in this text.

IDF represents the data of all text divided by the text containing the keyword, and then the obtained quotient is obtained by logarithm, that is, the frequency of the reverse word frequency is:

$$idf_i = \log \frac{|D|}{1+|\{j|t_i \in d_j\}|}$$

(6)

Where \(|D|\) refers to the total number of all texts, and the subheading refers to the total number of texts containing the keyword \(t_i\). In general, we need to add 1 to the denominator to prevent the word from being in the expected library, resulting in a dividend of 0.

So the formula for TF-IDF is:

$$tf_{i, idf} = tf_{i} \times idf_i$$

(7)

The idea of the TF-IDF algorithm is that if a word has a higher TF value in a text and fewer occurrences in other text, this keyword has a better ability to distinguish between different texts.
5.2 Principle of cosine distance

After obtaining the characteristics of words through TF-IDF, we use the cosine distance formula to compare the similarity between different entities.

The core idea of cosine similarity is the cosine value between the angles of two vectors in a vector space as a measure of the two individuals. The magnitude of the difference between them, the cosine value is close to 1, the included angle tends to 0, indicating that the two vectors are more similar, the cosine value is close to 0, and the included angle tends to 90 degrees, indicating that the two vectors are not similar. The formula is:

\[
\cos(\theta) = \frac{\sum_{i=1}^{n} (x_i \cdot y_i)}{\sqrt{\sum_{i=1}^{n} (x_i)^2} \cdot \sqrt{\sum_{i=1}^{n} (y_i)^2}}
\] (8)

Where \(x_i\) and \(y_i\) are the TF-IDF values in the two texts, respectively. We set a threshold of 0.9, that is, when the cosine distance is greater than 0.9, we assume that the two texts are the same entity.

6. Knowledge question answering application

This module is mainly composed of three parts: problem understanding, problem solving and answer generation. Among them, the problem understanding analyzes the input user problem, extracts the semantic information necessary to solve the problem, and the problem solving converts the result of the problem understanding into the query of the graph database, and the final answer generation converts the query result into a natural language answer. Return to the user. Because the intelligent question and answer based on the knowledge graph is conducted on the basis of the knowledge graph, the problem solving is finally reflected as a query on the knowledge graph, so the key of the problem understanding module is to graph the user question to the entity on the knowledge graph and express the user's query intent constitutes the semantic information to be extracted from the question. The overall flow of the module is shown in Figure 4.

Figure 4 Design of knowledge acquisition module

User questions are firstly segmented through shallow grammatical analysis, and the results obtained are used as input to the problem classification; after the problem type is identified, some user intentions of the problem have been identified, and the reference to the knowledge graph in the problem is also extracted. Refers to being linked to a specific entity in the knowledge graph; the result of the question understanding instantiates the graph database query template to generate specific query statements, and the resulting query results are processed into natural language answers by the answer generation template and returned to the user.

6.1 Problem understanding

Since user questions are directed to entities in the knowledge graph, user questions can always be split according to elements in the knowledge graph. Through analysis, we divide the user questions that need to be answered into several categories, such as: asking the maximum value of the attribute of a certain type of thing in a certain aspect, and judging whether there is a relationship between one instance and
another instance. User questions are first analyzed by shallow syntax to generate the results of word segmentation and part-of-speech tagging in sentences, and these results are input into the question template. For each type of problem, we have implemented a corresponding problem template to describe the characteristics of this type of problem, used to identify this type of problem, and at the same time extract the reference term for the entity in the knowledge graph in the problem, that is, named entity. For problem classification, machine learning methods are also used, such as: SVM, neural network and other methods.

Because the user does not understand the specific names of the entities in the knowledge graph, the named entities in the sentence may be literally inconsistent with the named entities extracted from the sentence. Therefore, we link the named entities to the corresponding entities in the knowledge graph through the similarity measurement method between the strings. When the similarity between the two candidate entities and the named entity in the sentence is the same, the simple similarity measure can no longer uniquely determine the corresponding entity. For this purpose, we achieve the purpose of disambiguation through user interaction. When an entity cannot be uniquely determined, the user will be asked to choose from the candidate entities.

6.2 Problem solving and answer generation
For each type of question, there are corresponding query and answer generation methods. The query process judges a series of entities output by the question understanding and generates a knowledge graph query corresponding to the user question. The query results are filled into the answer generation process corresponding to each type of question to generate answers that ultimately form in natural language. Based on the complexity of the answers generated, the questions are divided into three categories: The first category is the question that can be directly used as the answer, which is called simple question. The second category is based on the results of the query, through logical reasoning to get answers. The third category is based on the query results, and the answers are obtained through uncertain reasoning.

6.3 Example of knowledge acquisition
Enter the topic--keywords, and return to the article title according to the semantic similarity. Click on the title to jump to the article:

![Figure 5 Example of knowledge acquisition in the field of agriculture 1](image)

As shown in Figure 5, Match the query content with semantic similarity, and finally return the content of the article with the highest similarity-the abstract, to the user, and display the keywords in red. The result shown in Figure 6:
7. Summary
This paper mainly puts forward a complete system of knowledge graph in the agricultural field from scratch to use. The new generation of search engine technology has important significance for the theoretical research and practical application value in the agricultural field.

However, with the in-depth study of knowledge graphs, knowledge graphs have quality problems such as incompleteness and wrong relationships in various [7]. These problems have great impact on applications such as intelligent question answering. Knowledge reasoning is an important means of knowledge graph completion. It can use the existing explicit knowledge in the knowledge graph to predict the implicit knowledge that has not been stored in the graph, and gradually complete the knowledge graph. Therefore, it has become a research hotspot of knowledge graph. First, the neural network-based knowledge reasoning method has stronger reasoning ability and generalization ability, higher utilization rate of entities, attributes, relationships and text information in the knowledge base, and better reasoning effect[8].

In addition, the goal of artificial intelligence is to enable computers to become as intelligent as humans. How to help people make decisions is a great value of their applications. I hope this article can provide some help for the acquisition of knowledge in the field of agriculture.

References
[1] Bin Ge,Zhen Tan. Construction technology of military knowledge graph [J]. Journal of command and control,2016,2(4):302-308
[2] Huilin Tang. Research on knowledge graph completion algorithm based on structure and semantic information fusion [R]. Beijing University of Posts and Telecommunications,2016
[3] Tong Yao,Mengjie Wang,Haofen Wu. Research on the construction and application of vertical knowledge graph [J]. Knowledge Management Forum,2016,1(3):226-234
[4] Qiao Liu,Yang Li,Hong Duan. A survey of knowledge graph construction technology [J]. University of Electronic Science and technology,2015
[5] Gannis Bekoulis,Johannes Deleu.Joint entity recognition and relation extraction as a multi-head selection problem[J]. Ghent University,2018
[6] AUER S, BIZER C, KOBILAROV G, et al. Dbpedia: A nucleus for a web of open data[M].Berlin:Springer Berlin Heidelberg, 2007.
[7] PAULHEIM H, PONZETTO S P. Extending DBpedia with Wikipedia List Page[C]//NLP-DBPEDIA@KWC, 2013
[8] Chen Xueqi,Jin Xiaolong,Wang Yuanzhuo,et al.Survey on big data system and analytic technology[J].Journal of Software.2014.25(9):1889-1908(in Chinese)