ISBN: 978-981-94-1156-6
2024 the 14th International Workshop on Computer Science and Engineering (WCSE 2024)
doi: 10.18178/wcse.2024.06.019

Research on the Construction of Knowledge Graph of Strawberry Diseases and Pests

Long Zhao[1,2], Yin Xu[2], Yanyan Wang[2 +], Fei Li[1,2] and Qiangzhong Feng[2]

[1.]School of Computer Science and Technology, University of Science and Technology of China, Hefei 230022, China

[2.]Innovation + Research Institute, GuoChuang Cloud Technology Ltd, Hefei 230031, China

Abstract. To solve the problems that existing knowledge graph extraction methods in the field of pests and diseases require a large amount of annotated data, and the trained model has limited scalability and versatility, this paper proposes an unsupervised graph data extraction paradigm based on the Universal Information Extraction (UIE) model. According to the characteristics of the pest and disease corpus, the data processing mode is designed, while the relationship and attribute types are predefined. UIE model is used for extracting tail entities. And the attribute values are matched for content based on semantic similarity. By fusing multi-source information, a strawberry pest and disease knowledge graph is constructed. A total of 427,304 triples are extracted in this study, and the F1-score of the extraction results reached 87.2%. Compared with the existing methods, the granularity of extraction is more detailed, with strong scalability, and there is no need to consume resources for additional data annotation. The knowledge graph constructed by this research can subsequently provide a high-quality knowledge base for strawberry disease and pest knowledge question and answer systems, intelligent recommendation systems and other downstream applications.

Keywords: knowledge graph, knowledge extraction, strawberry disease and pest, information extraction

1. Introduction

Strawberries' cultivation area and production have been increasing yearly. However, they are easily affected by pests and diseases, which mainly affect their yield and quality. Traditional pest control methods based on experience and manual assessments are inefficient and prone to errors. Although the internet is widely used, search engines provide a lot of web pages, but users still need to search through them to find relevant information. This scattered and redundant data distribution is time-consuming and laborious, limiting the rapid and efficient progress of agricultural informatization.

The concept of Knowledge Graph (KG) was proposed by Google in 2012, which is essentially a structured knowledge representation that organizes knowledge in the form of graph [1]. Current knowledge graphs in the open domain are based on large-scale commonsense knowledge, such as YAGO [2], DBpedia [3], Freebase [4], WordNet [5], and so on. This form of knowledge representation is more efficient and concise, so it is of great practical significance and application value to construct a knowledge graph of strawberry pests and diseases. In this study, we propose an unsupervised agricultural knowledge graph construction method, which performs knowledge triples extraction by designing a data processing approach and predefining the types of relational attributes, as well as similarity calculation. The knowledge graph provides rich information on pests and diseases characteristics and transmission pathways, which can be used to detect and control the occurrence and spread of pests and diseases in a timely manner. By analyzing the data and patterns in the knowledge graph, it can provide strong support for the development and application of pest control technologies, and improve the efficiency of strawberry production.

In summary, the contribution of the work in this paper is as follows: (1) A knowledge graph of strawberry pests and diseases was constructed, containing 65,075 entities, 61 relationships, and 427,304 triples. (2) An unsupervised mapping data extraction paradigm based on an information extraction model is proposed for

+ Corresponding author.
E-mail address: wang.yanyan@ustcinfo.com.

extracting fine-grained knowledge triples from pests and diseases corpus texts without the need to annotate the texts.

2. Related Work

In recent years, with the rapid development of knowledge graph research, knowledge graph research in agriculture has also gradually attracted the attention of researchers. Wang [6] constructed a tobacco disease knowledge graph using a top-down approach. Qiao [7] derived knowledge graph hierarchy and data level transformations from narrative lists. Zhu [8] researched recommender systems by modeling user interests and using wheat pest and disease data. Zhang [9] built an apple pest and disease knowledge graph from web crawlers and specialized books, improving application accuracy. Zhang et al. [10] created a fine-grained apple pest and disease knowledge graph, providing intelligent assisted diagnosis and other functions. Callahan et al. [11] explores the use of open-source LLMs for the creation of ontologies and knowledge graphs.

Current agricultural knowledge graph research heavily relies on labelled data trained with supervised models. Label differentiation is challenging due to diverse entities and relationships in ontology construction. This annotation process is laborious and may limit models' scalability and generalization. Models for specific crops' pests and diseases may not generalize well to other crops. To address these issues, we propose an unsupervised unified atlas construction paradigm based on an information extraction model. This involves organizing and summarizing data, setting multiple extraction modes for efficient and accurate knowledge extraction of different text categories, and storing the extracted information as triples.

3. Strawberry Pest and Disease Knowledge Graph Construction

3.1. Data Processing

This study involves two main data types: articles on strawberry pests and diseases obtained from agricultural websites, and a pesticide information table containing details on common pesticides. Since the strawberry pests and diseases articles share a similar structure, we organized the text into a unified semi-structured format using specified rules to simplify model extraction and improve accuracy. This format typically includes basic pest and disease details (aliases, English names, onset temperatures), prevention and control methods (chemical, agricultural), and infestation sites.

In this study, we process various information categories by labeling text, facilitating the model's extraction of detailed data. As articles primarily focus on the same disease, it's treated as a unified head entity alongside pathogen-related info. To precisely depict relationships among entities, we predefined relationship and attribute sets for diseases and pests based on data content, business needs, and domain expert guidance. Diseases have 25 predefined relationship types, pests have 26, and both have 5 attribute types. The details are shown in Table 1:

Table 1: Predefined relationships and attributes

Classification	Caption	Predefined Relationships	Predefined Attributes
Diseases	Synopsis	English name, Alias, Classification, Site, Crop, Distribution area, Temperature, Humidity, Period, Pathogens	——
	Pathogens	English name, Shape, Size, Suitable temperature, Suitable humidity, Suitable area	——
	Systematics	Class, Order, Family, Phylum, Subphylum, Species	——
	Agricultural prevention and control	Resistant varieties	Agricultural control methods
	Chemical control	Pesticides	Chemical control methods
	Transmission route	Transmission factors	Mode of transmission
	Symptomatic	——	Symptomatic
	Effect	——	Effect

Pests	synopsis	Alias, Scientific name, Species, Crop, Infected positions, Distribution area, Suitable temperature, Peak time	——
	Systematics	Class, Order, Family, Species, Phylum, Subphylum	——
	Exterior condition	Adult body length, Adult body width, Adult color, Adult shape, Larval body length, Larval body width, Larval color, Larval shape	——
	Chemical control	Pesticides	Chemical control methods
	Biological control	Biopharmaceuticals, Natural enemies	Biological control methods
	Physical control	Tools	Physical control methods
	Symptomatic	——	Symptomatic
	Effect	——	Effect

3.2. Ontology Construction

An ontology is an explicit specification of a conceptual model [12], and by constructing an ontology for strawberry pests and diseases, it is possible to be more explicit about what is being extracted from the text and how the extraction model is constructed. The strawberry pest ontology was controlled into two layers, including eight categories of parent concepts, namely crop, taxonomy, name, appearance, environment, pathogen, pest, and control methods.

3.3. Knowledge Graph Construction

Knowledge graph construction is divided into top-down and bottom-up approaches. This study combines the two approaches, firstly determining part of the model information through industry general knowledge as well as expert experience, then summarizing and organizing according to the data content, and finally instantiating the constructed ontology structure through information extraction. The specific construction process is shown in Fig.1:

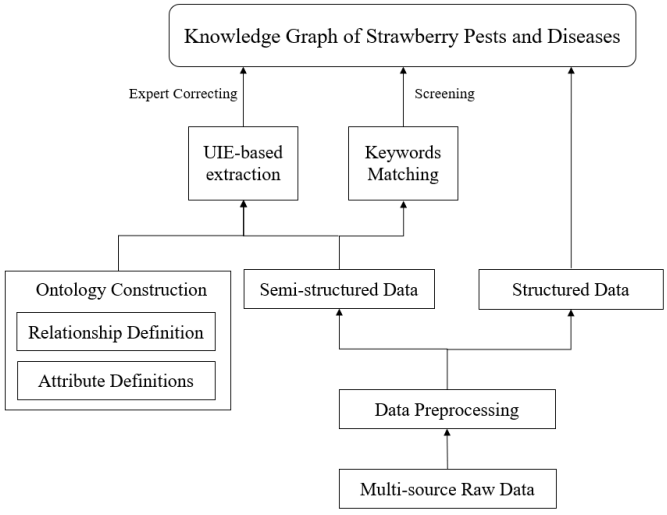

Fig. 1: Process for Building a Knowledge Map of Strawberry Diseases and Pests

UIE-based Extraction. UIE [13] is a unified information extraction model that simplifies different types of extraction tasks using a text-to-structure approach. In this study, we use a predefined strawberry pest and disease relationship type for extraction, leveraging the UIE model. The extracted content corresponds to the tail entity of the defined relationship. As shown in Fig.2, in the text related to strawberry powdery mildew, "*strawberry powdery mildew*" is the main descriptive object of the subsequent text, so it can be directly used as the co-head entity of the relationship-tail entity extracted later. For the paragraph titled "*Introduction*", the predefined extraction modes are "*Alias*", "*English Name*", etc. The UIE model is used to extract the predefined patterns. The UIE model is called to extract the predefined contents and construct the information into triples data for subsequent knowledge graph construction.

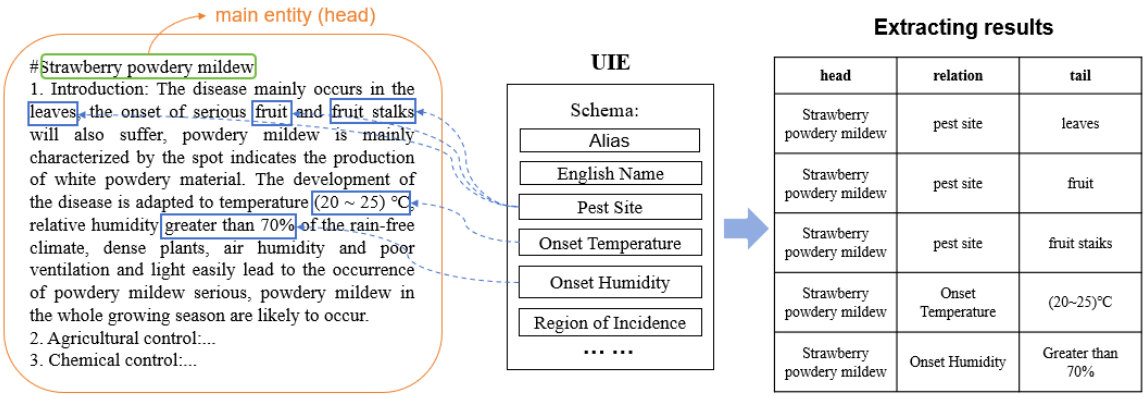

Fig. 2: Example of triple extraction

Attributes re-extraction. For longer descriptive text information, such as specific control methods, onset symptoms, etc., they can either be stored as attributes or reextracted to enrich the content of the map as well as to show the commonalities between different pests and diseases. The number of keywords summarizing $\boldsymbol{W}_i = \{\boldsymbol{w}_i^1,...,\boldsymbol{w}_i^k\}, k$ as attributes for attribute i based on the keywords of agricultural experts for attributes for attribute category i is used to use the pre-trained language model BERT [14]. Word embedding is performed for all keywords to obtain the embedding set of keywords $\boldsymbol{E}_i = \{\boldsymbol{e}_i^1,...,\boldsymbol{e}_i^k\}$. The attributes text a_i^m corresponding to attribute i is sliced and a context-based embedding representation $\{\boldsymbol{a}_i^1,...,\boldsymbol{a}_i^j\}$ is generated using BERT, j is the number of sliced texts. The cosine similarity between the text and the keywords is calculated as the text similarity, where the cosine similarity between the text a_i^m and the keyword $\boldsymbol{e}_i^m$ is calculated as shown in Equation (1), and finally the text is matched to the keyword that has the maximum similarity with it.

$$Similarity(\boldsymbol{a}_i^m, \boldsymbol{e}_i^m) = \frac{\boldsymbol{a}_i^m \cdot \boldsymbol{e}_i^m}{\left\|\boldsymbol{a}_i^m\right\| \times \left\|\boldsymbol{e}_i^m\right\|} \tag{1}$$

As shown in the example in Fig.3, for agricultural control measures, the content can be summarized and generalized, and then the attributes text can be segmented, and each segment of text can be matched and corresponded to the generalized content by calculating the text similarity to achieve the expansion of the mapping data.

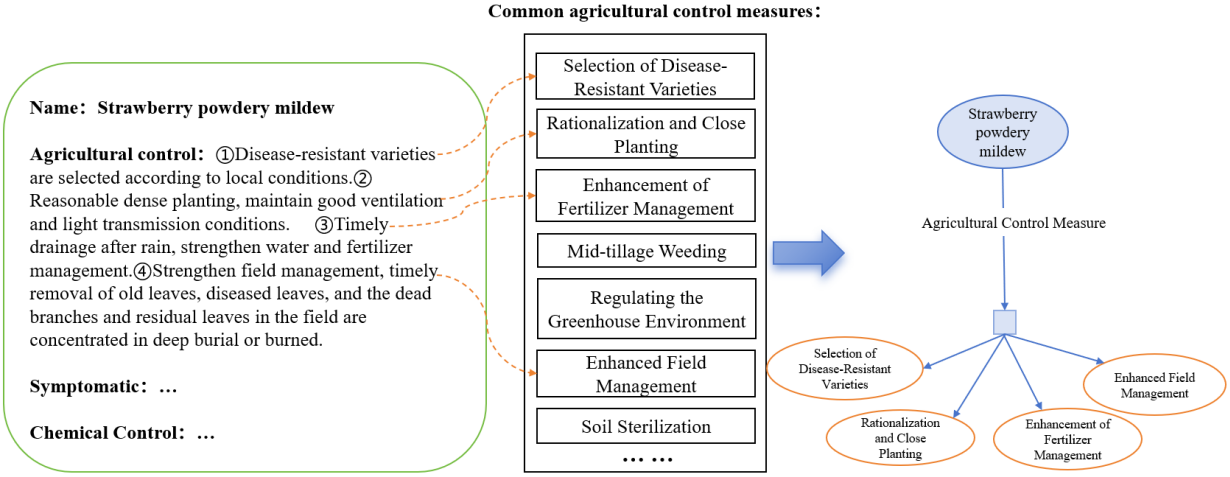

Fig. 3: Example of re-extracting attributes

Knowledge Fusion. Articles may refer to the same entity differently, such as "*leaf*", "*leaf blade*", and "*foliage*" for the same plant part in the atlas. To harmonize these variations, we standardize strawberry organ names and match data using similarity metrics. Since our data comes from varying-quality web crawlers, we analyze multiple articles on the same pest or disease. Credible information must appear in at least 30% of articles. For instance, "*strawberry anthracnose-control pesticide-pyrimethanil*" was extracted from 60% of articles on strawberry anthracnose, making it eligible for our strawberry pest and disease knowledge map.

4. Construction Results

To evaluate the advantages and disadvantages of the extraction results, we adopt three commonly used evaluation metrics in the field of information extraction, which are $\boldsymbol{Precision}$, $\boldsymbol{recall}$ and $\boldsymbol{F1\_score}$. However, to the best of our knowledge, there is currently no unsupervised approach for knowledge graph construction. Therefore, we employ the method of prompting LLMs as the comparative baselines. We utilize ERNIE-Bot v3.5 and SparkDesk v3.5 for triplet extraction.

Randomly extract 100 paragraphs of text for manual labeling, and calculate the extraction results of the indicators shown in Table 2. Significantly, our method achieves the best performance in all metrics. The $F1\_score$ of the extraction results of the method proposed in this paper is 87.2%, which indicates that this method not only saves time and labor, but also the extraction results are more credible, without a lot of manual correction.

Table 2: Evaluation of extraction results

Method	$Precision$/100%	$Recall$/100%	$F1\_score$/100%
Prompt LLM (ERNIE-Bot v3.5)	48.9	18.9	27.5
Prompt LLM (SparkDesk v3.5)	50.0	21.8	30.3
Ours	**84.6**	**89.9**	**87.2**

After the final de-duplication and fusion process of the triples from different data sources, the statistical results are shown in Table 3:

Table 3: Result statistics

Entities	Relations	Triples
65,075	61	427,304

5. Conclusion

By analyzing the present state of agricultural knowledge graph construction, we introduce an unsupervised unified graph construction paradigm rooted in an information extraction model. This addresses the challenging task of labeling numerous samples, which is often tedious, time-consuming, and labor-intensive. We apply this paradigm to building a strawberry pests and diseases knowledge graph, extracting ternary data, and then performing entity alignment and knowledge fusion by computing semantic similarity.

With strawberry pest and disease research progressing, knowledge must be regularly updated. This necessitates continuous extraction of knowledge from recent research to construct a more comprehensive knowledge graph. These graphs are invaluable for tasks like quizzes and intelligent recommendations. Our created graph can be leveraged to build a Q&A system that answers farmers' queries, offering technical guidance. This empowers farmers to maintain and enhance their yields, ultimately leading to increased income.

6. References

[1] A. Carlson, J. Betteridge, B. Kisiel, et al. Toward an architecture for never-ending language learning. In: A. Carlson, et al. *Proc. of the Twenty-Fourth AAAI Conference on Artificial Intelligence.* 2010, 24 (1): 1306-1313.

[2] F.M. Suchanek, G. Kasneci, G. Weikum. AGO: A Core of Semantic Knowledge Unifying WordNet and Wikipedia. In: F.M. Suchanek, et al. *Proc. of the 16th international conference on World Wide Web.* 2007, pp.697-706.

[3] S. Auer, C. Bizer, G. Kobilarov, et al. DBpedia: A Nucleus for a Web of Open Data. In: S. Auer, et al. *Proc. of International semantic web conference.* 2007, pp.722-735.

[4] K.D. Bollacker, C. Evans, P. Paritosh, et al. Freebase: a collaboratively created graph database for structuring human knowledge. In: K.D. Bollacker et al. *Proc. of the 2008 ACM SIGMOD international conference on Management of data*, 2008, pp.1247-1250.

[5] G.A. Miller. WordNet: a lexical database for English. In: G.A. Miller, *Proc of Human Language Technology. Association for Computational Linguistics*. 1995, 38(11): 39-41.

[6] J. Wang. Research on tobacco disease prevention and control model based on case-based reasoning and knowledge graph, 2016.

[7] B. Qiao. Research on Knowledge Graph Construction Technology Based on Agricultural Narrative Lists, 2019.

[8] Y. Zhu. Research on recommendation technology based on knowledge graph of wheat pests and diseases, 2022.

[9] H. Zhang. Research and development of apple disease identification and application assisted decision-making system based on deep learning, 2021.

[10] J. Zhang, M. Guo, Y. Zhang et al. Research on the construction of fine-grained apple pest and disease knowledge graph. In: J. Zhang, et al. *Proc. of Computer Engineering and Applications*. 2023, 59(05): 270-280.

[11] T. Callahan, I. Tripodi et al. An open-source knowledge graph ecosystem for the life sciences. In: T. Callahan et al. *Proc. of Scientific Data*. 2024, 11.1: 363.

[12] T.R. Gruber. A translation approach to portable ontology specifications. In: T.R. Gruber. *Proc. of Knowledge Acquisition*. 1993, 5(2):199-220.

[13] Y. Lu, Q. Liu, D. Dai, et al. Unified Structure Generation for Universal Information Extraction. In: Y. Lu, et al. *Proc. of the 60th Annual Meeting of the Association for Computational Linguistics*. 2022, pp.5755-5772.

[14] J. Devlin, M.W. Chang, K. Lee, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: J. Devlin, et al. *Proc. of NAACL-HLT*. 2019, pp.4171-4186.