Cn-MAKG: China Meteorology and Agriculture Knowledge Graph Construction Based on Semi-structured Data

Qi Chenglin, Song Qing, Zhang Pengzhou[*], Yuan Hui
New Media Institute
Communication University of China
Beijing, China
qclsilence@gmail.com, songqing@cuc.edu.cn, zhangpengzhou@cuc.edu.cn

Abstract—**In this paper, a method of constructing China meteorology and agriculture knowledge graph based on semi-structured data is proposed. Firstly, demand analysis, determine the boundary of knowledge, according to which design the schema of Cn-MAKG. Then the semi-structured knowledge data is preprocessed in accordance with the standard indexes in the domain of meteorology and agriculture. Finally, the graph database Neo4j is used to store the knowledge graph, so as to realize the construction of Cn-MAKG. At present, the knowledge graph has been successfully applied to the automatic generation of crop meteorological reports.**

Keywords—meteorology and agriculture; knowledge graph; schema; semi-structured data; data processing; knowledge storage;

I. Introduction

China has a vast territory, different geographical regions in China have the different meteorological performance at different time periods. As the same time, China is also a big agricultural country with many kinds of crops. The planting type and growth of these crops are affected by geographical regions and meteorological conditions. In order to protect and promote agricultural production, it is necessary to consider how to give full play to the functions of meteorology in agricultural production. The agrometeorological department is responsible for the weather forecast, agrometeorological observation and prediction, the assessment of agrometeorological disasters and related ten-day reports or monthly reports publication for governmental decision-making departments and crop growers. These services are dependent on the knowledge and data related to meteorology and agriculture such as Agro climate geographical areas division in China, meteorological factors in these areas, the influence of meteorological factors on different crops in different growth periods and the observed data from different meteorological stations at different time periods.

In order to provide better services and promote agricultural production, we must explore how to use information science and technology to arrange and represent agrometeorological knowledge, collect and store meteorology data in different regions at different time periods, analyse these data combined with the domain knowledge, and finally generate crop meteorological reports automatically. With the development of modern AI, knowledge graph is introduced into the field of knowledge engineering, which is used to describe the Relations between entities and construct knowledge hierarchy in a specific or open domain knowledge.

Combining with the schema design, knowledge acquisition, knowledge storage and other related technologies, this paper proposes and realizes a construction method of China meteorology and agriculture knowledge graph, and it may be applied to the automatic generation of meteorological and agricultural related texts.

II. The theory of knowledge graph

A. Knowledge Graph

Knowledge graph is a kind of semantic network or a graph-structured knowledge base. It includes three key elements: nodes, attributes (properties) and Relations. The convention for knowledge graph is to use nodes to represent concepts and entities in the word, and use edges to represent the Relations between nodes. As a new way of knowledge representation, knowledge graph is becoming an important branch of the framework of artificial intelligence, and share a strong historical bond with semantic web, ontology and so on.

One of the original ways of knowledge representation was the Semantic Network proposed by M. Ross Quillian and Robert F. Simmons in 1968. [1] In Semantic Networks, entities represent classes, concepts and so on, while edges represent the Relations between them, mainly *is-a* and *part-of*. However, Semantic Networks have no formal syntax and semantics, nor does it allow users to customize labels on nodes and edges. In 1989, Tim Berners Lee, the father of the World Wide Web, proposed the Semantic Web. [2] Semantic web is no longer just building links between web pages, but connecting all the resources on the network, and the types of Relations between

[*] Corresponding author

978-1-5386-5892-5/18/$31.00 ©2018 IEEE
ICIS 2018, June 6-8, 2018, Singapore

them can be customized. Using RDF, OWL or other W3C standards to describe resources, and defining uniform resource identifiers for resources, which can enable data from different domains to link and share, and these data are called Linked Data. [3] At present the Linked Open Data(LOD) project has brought much high-quality knowledge based on Wikipedia together, such as DBpedia[4], Yago[5], Freebase[6] and so on. These knowledge bases have laid the foundation for the success of Google knowledge graph. Since Google launched the "knowledge graph"[7] in May 2012, the related research of knowledge graph has rapidly raised the upsurge in the domestic industry and academia.

B. *Knowledge graph construction technologies overview*

The logical architecture of knowledge graph is divided into two layers. At the top is schema layer, and the bottom layer is data layer. There are usually two approaches to construction a knowledge graph, one is to start with designing schema layer, and then fill in the data of specific instance (entity), which is called top-down approach. Another one is to extract entities from mass data collected first, and then induce and abstract these entities to form an ontology concept hierarchy as the schema layer, which is called bottom-up approach. In the actual construction process of knowledge graph, these two approaches usually mixed. It's more important to constantly adjust, enrich and improve the knowledge graph we have constructed according to the existing data and actual requirements.

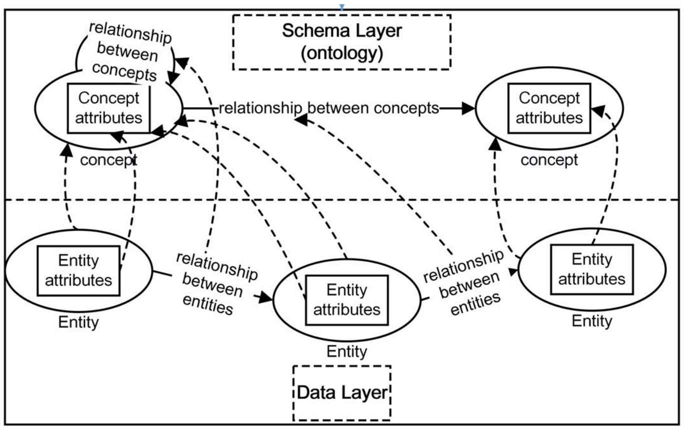

Fig. 1. The logical architecture of knowledge graph

The construction of knowledge graph is generally divided into three processes: knowledge acquisition, knowledge fusion, knowledge calculation and further processing.

Knowledge acquisition needs to consider the source of knowledge and the structure of data. Consideration of knowledge sources mainly refers to ascertain whether knowledge is single or multi-source before the graph is constructed, which determines whether the work of knowledge fusion is needed, although sometimes need to disambiguate entity mentions when the knowledge is from a single source. In addition, analyzing the data is unstructured, semi-structured or structured is needed, which determines the technical means adopted in the knowledge acquisition phase. Knowledge acquisition mainly refers to the acquisition of entities, attributes, and Relations. If the data is unstructured like text, it is necessary to use the methods of Natural Language Processing like named entity recognition (NER), Relation extraction, and some basic techniques of NLP such as segmentation, the dependency parsing and so on. If the data is semi-structured data from web pages, it's necessary to crawl data, parse them, and then transform them into structured data.

The target of knowledge fusion is to integrate the knowledge of different sources into a large knowledge base. The tasks of knowledge fusion include schemas fusion, entity disambiguation (or entity linking), and attributes fusion.

For the sake of constructing a high quality, stable and continuously updated knowledge graph, knowledge calculation and further processing are very important. This work includes evaluation of knowledge graph, implicit information supplement by knowledge reasoning and the update of knowledge graph.

Our knowledge graph *Cn-MAKG* in this paper is based on semi-structured, so the technologies we have involved include schema design, knowledge acquisition, data preprocessing and knowledge storage.

III. Construction of Cn-MAKG

A. *Schema design of Cn-MAKG*

Before constructing our knowledge graph, we have to do the schema design first. The schema design of the knowledge graph refers to the ontology design. Before designing, we have to consider the actual requirements and application targets, in order to determine the boundary of knowledge needed for knowledge graph construction. Because the knowledge graph in this paper describes the connections between crop planting in different regions of China and the meteorological conditions in local different time periods, we need to research on related domain knowledge about the division of Agro climate geographical areas in China, the types of crops planted in different Agro climate geographical areas and the growth habits of these crops in different regions and time. The research on the knowledge is mainly based on the "Manual of meteorological service for staple crops" edited by China Meteorological Press and Liuxi Mao, director of National Agrometeorological Center of China. [17]

The schema of the knowledge graph in this paper includes three part of concepts related China's geography, crops and meteorology. The concepts related to geography include countries, Agro climate geographical areas and its subareas, the provinces (include province-level municipalities) in these subareas. The concepts related to crops include crop types, the growth period of the different crop, the beginning and ending time of growth period. The concepts related to meteorology include effects of meteorology conditions on crops and climate background. Among them, each concept has their corresponding attributes.

The semantic Relation of the schema in knowledge graph refers to the Relations between concepts. Semantic Relations can be imposed constraints by setting some attributes. In our Cn-MAKG, some semantic Relations have some attributes as constraint conditions. All semantic Relation types in our knowledge graph are as follows:

TABLE I. RELATION TYPES IN CN-MAKG

Custom Semantic Relation Name	**Explanations Of Semantic Relations**
BELONG_TO	The Relation between instance and concept, or the Relation between the part and the whole. For example, the Relation between of “China” (instance) and “Country” (concept), “JiangNan District” (part) and “China” (whole).
IS_A	The Relation between subclass and class. For example, the Relation between “single-season rice in JiangHan District” (subclass) and “single-season Rice” (class).
PLANT	The Relation between Agro climate geographical areas and the crops planted in these areas. For example, the Relation between “JiangHuai District” and “winter wheat”.
GROWTH_PERIOD_IS	The Relation between crops and growth periods. For example, the Relation between “rape” and “5-leaf stage”.
POSITIVE_METEOROLOGY_CONDITION	The Relation between growth periods and the positive effect results of meteorology. For example, the Relation between “tillering stage of single-season rice in JiangHan District” and “Good for growing”. In this Relation, there are some meteorology indexes to transform into the attributes and their value. For example, a triple of a Relation and one of its attributes is (POSITIVE_METEOROLOGY_CONDITION, daily mean temperature, [15°C,18°C]).
NEGTIVE_ METEOROLOGY_ CONDITION	The Relation between growth periods and the negative effect results of meteorology. For example, the Relation between “tillering stage of single-season rice in JiangHan District” and “The stem is fragile and causing rice blast”. In this Relation, there are also some meteorology indexes to transform into the attributes and their value.
TIME_IS	The Relation between growth and their start and end time. For example, the Relation between “Seeding stage of spring corn in HUABEI District” and “April 21st to May 10th”.
CLIMATE_BACKGROUND_IS	The Relation between growth periods and climate background. For example, “Seeding stage of spring corn in HUABEI District” and “average sunshine time is 7~10 hours”.

The schema of China meteorology and agriculture knowledge graph in this paper is designed as follows:

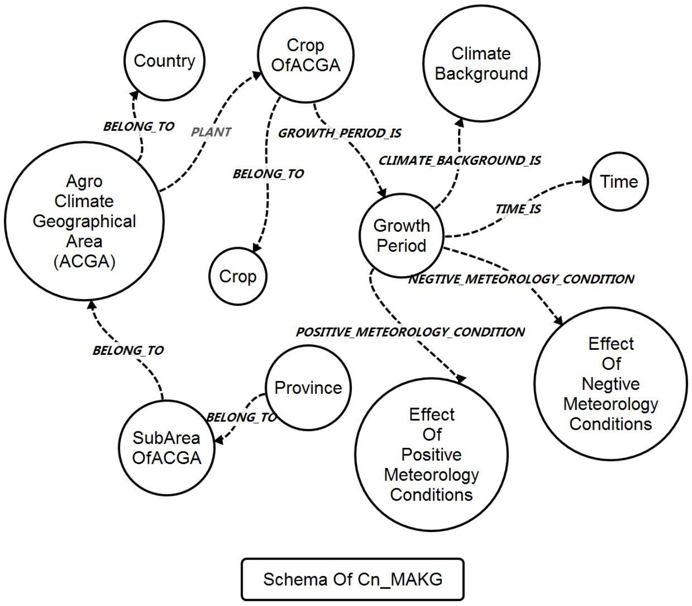

Fig. 2. The schema of Cn-MAKG

B. *Knowledge acquisition of Cn-MAKG*

Knowledge acquisition of knowledge graph includes the acquisition of entities, relations, and attributes of them.

Firstly, what should be considered is the problem of knowledge source. After the investigation, we conclude that some professional data such as Agro climate geographical areas division in China, meteorological factors in these areas, the influence of meteorological factors on different crops in different growth periods are either rarely open, or the knowledge standards are different, or quality of data is too poor. In order to overcome these problems, we choose the professional and authoritative book “Manual of meteorological service for staple crops” as the source of knowledge and data.

As for the structure of knowledge data, the data in this book are tabular, but not strictly structured data, thus it is necessary to transform these semi-structured data into structured data.

For constructing the knowledge graph in this paper, 10 major cropping crops in China are selected as the research object, and we divide China into 17 Agro climate geographical areas according to the division standard in the book “Manual of meteorological service for staple crops”. [17] The knowledge we need include phenology calendar of crops, meteorological conditions, the indexes during the growth periods of crops and so on. In order to get these knowledge data on paper books, we use OCR scanning software to computerize these semi-structured table data.

C. *Data preprocessing of Cn-MAKG*

The semi-structured data on the book exists several problems. First, the recognition accuracy of OCR software has not reached 100%, so there are some errors in the data. Second, some of the data in the table on the book are not standard forms, such as *Time* data. Moreover, some textual data related to meteorological knowledge are not well quantified according to the meteorological index. Respond to the above-mentioned problems, carrying on the data in pre-process measurement is necessary. In this paper, we need to clean and transform the electrical data including correcting the data not correctly recognized, normalizing the data in a non-standard format, quantifying some text descriptions according to meteorological indexes we summarized, and realizing the standard expression of these quantitative conclusions by using custom rules.

All tasks of data preprocessing for constructing our knowledge graph are as follow (note that the data we need to preprocess are Chinese):

TABLE II. TASKS OF DATA PREPROCESSING

Task Of Data Preprocessing	**Task Example**
error data correction	eg. “日平均气温” is misrecognized as “曰平均气温”
data normalization	eg. “下／9～中／10” is normalized as “09 月 21 日-10 月 20 日”
text descriptions quantification	eg. “阴雨日数达 3 天以上” is quantified as “日降水量大于等于 0.1 毫米” ∩ “日均日照少于 3 小时” ∩“降水量持续天数大于等于 3 天”
quantitative conclusions standardization	eg. the attribute “日平均气温 15～18°C” of Relation “positive_meteorology_condition” is expressed as “日平均气温:[15°C,18°C]/有利于生长” according to custom rules; The negative meteorology_condition and the negative result “阴雨寡照天气，不利开花结荚” is quantified and expressed as “日降水量:>=0.1 毫米+日均日照:<=3 小时/不利培育壮秧/阴雨寡照”

After completing the preprocessing tasks, the data need to be parsed and extracted according to the custom rules, and the results are finally integrated into a table. The table fields are designed as follows:

TABLE III. FIELDS IN THE TABLE OF DATA PREPROCESSING RESULTS

Field
CropType
GrowthPeriod
MainProductionArea
ProvinceInMainProductionArea
Time
PositiveMeteorologyCondition
NewPositiveMeteorologyCondition
NegtiveMeteorologyCondition
NewNegtiveMeteorologyCondition
ClimateBackground

D. *Knowledge Storage of Cn-MAKG*

At present, the standard ways to store knowledge graph are using RDF (namely, triples storage) and using the graph database. In this paper, we select the latter and we select the most popular graph database Neo4j, which is available in a GPL3-licensed open-source "community edition", has excellent and complete query language Cypher, and supports all kinds of graph algorithms, to store knowledge graph. There are three ways to construct knowledge graph using Neo4j: the first one is to use the native API provided by Neo4j to complete the construction by coding according to the designed schema. The second one is to use Object Graph Mapping (OGM) to construct. And the other one is to assemble the data into node or node-Relation CSV files according to the schema, then use some tools to import these CSV files into neo4j.

In this paper, we choose the third manner. There are several ways to import CSV files of nodes and Relations into Neo4j including directly using the “LOAD CSV” query statement of Cypher, “Batch Inserter” in Java API provided by Neo4j, “Batch Import” tool based on “Batch Inserter”, and the official tool “Neo4j-import”. These methods have their own advantages and disadvantages and their application scenarios. The knowledge graph data was imported by using the official tool “Neo4j-import” in this paper. The most important thing in this way is assembling the data into CSV files according to the designed schema. There are two kinds of CSV files to be assembled, one is node CSV files, another one is Relation CSV files. According to the official regulations of Neo4j, both of them have their corresponding header format.

The first column in a node CSV file is the label of the node, and the form is “l:label”. The second column is the index of the node, and the form is “id:string:indexName”. Other columns are used to store the attributes (or properties) of the node. The header of the node “GrowthPeriod” in our knowledge graph is as follow:

TABLE IV. THE HEADER OF GROWTHPERIOD.CSV

l:label	id:string:growth_period_id	name	growth_period_order	crop_flag

The first column in a Relation CSV file is the index of the start node in a triple, the second column is the index of the end node, the third column is the Relation type, and other columns are the attributes (or properties) of Relation. The header of the Relation between the node “GrowthPeriod” and the node “NegtiveEffectResult” is as follows:

TABLE V. THE HEADER OF REL_GROWTH_PERIOD_OF_ACGA_NEGTIVE_EFFECT.CSV

id:string:growth_period_of_acga_id	id:string:negtive_effect_result_id	rel_type	daily_mean_temperature	…

After completing the assembly of the CSV files, you can use the command of the Neo4j-import tool to import these nodes and Relations data into Neo4j.

The numbers of nodes and Relations in our knowledge graph are shown in the following table:

TABLE VI. THE NUMBERS OF NODES AND RELATIONS

Knowledge Graph Elements	**Numbers of Elements**
Node labels	12
Relation types	8
Property keys	25
Nodes	907
Relations	4382

IV. CONCLUSION

The purpose of constructing China meteorology and agriculture knowledge graph in this paper is to provide relevant knowledge for the automatic generation of crop meteorological reports. At present, the knowledge graph has been successfully applied to the automatic generation of crop meteorological reports. At the same time, we also use D3.js to visualize the query results of the entities and Relations in the knowledge graph. In addition to the successful applications mentioned above, the domain knowledge graph also has many applications in industry, such as question answering system, intelligent information retrieval and so on. In the academic research, many research achievements have been achieved in the construction of knowledge graph. Due to the different data structure of knowledge sources, there are different methods to build knowledge graph. This paper is an exploration and an attempt to construct knowledge graph based on semi-structured data. Although our knowledge graph has been preliminarily constructed, we need further research on evaluating, enriching and updating of it in the future.

At present, the construction of knowledge graph is still facing many problems, especially high-quality knowledge data missing. Therefore, in the future, we need further research and explore more methods of the constructing knowledge graph.

ACKNOWLEDGMENT

This work was supported by the project of National Key Technology Research and Development Program of Ministry of Science and Technology of China (2014BAK10B01), Communication University of China engineering planning project (3132014XNG1422).

REFERENCES

[1] F.J. Sowa, "Principles of Semantic Networks: Exploration in the Representation of Knowledge." Frame Problem in Artificial Intelligence 2-3, 1991, pp.135-157.

[2] T. Berners-Lee, J. Hendler, and O. Lassila. "The Semantic Web: A New Form of Web Content That is Meaningful to Computers Will Unleash a Revolution of New Possibilities." Scientific American 284.5, 2001, pp.34-43.

[3] B. Christian, H. Toml, and T. Berners-Lee. "Linked data: the story so far." International Journal on Semantic Web & Information Systems 52.3, 2011, pp.1-22.

[4] J. Lehmann. "DBpedia: A large-scale, multilingual knowledge base extracted from Wikipedia." Semantic Web 6.2, 2015, pp.167-195.

[5] FM. Suchanek, G. Kasneci, and G. Weikum. "Yago - A Large Ontology from Wikipedia and WordNet." Web Semantics Science Services & Agents on the World Wide Web 6.3, 2008, pp.203-217.

[6] K. Bollacker, R. Cook, and P. Tufts. "Freebase: A Shared Database of Structured General Human Knowledge. " AAAI Conference on Artificial Intelligence, July 22-26, 2007, Vancouver, British Columbia, Canada DBLP, 2007, pp.1962-1963.

[7] A. SINGHAL. "Introducing the knowledge graph: things,not strings." https://googleblog.blogspot.com/2012/05/introducing-knowledge-graphthings-not.html, May 2012.

[8] GL. Qi, H. Gao, and TX.Wu. "The Research Advances of Knowledge Graph." Technology Intelligence Engineering, 2017.

[9] Q. Liu, Y. Li, H. Duan, Y. Liu, and Z. Qin."Knowledge Graph Construction Techniques." Journal of Computer Research & Development, 2016.

[10] ZL. Xu, YP. Sheng, LR. He, and YF. Wang. " Review on knowledge graph techniques." Journal of University of Electronic Science and Technology of China, vol.47, 2016, pp.589-606

[11] JZ. Pan, G. Vetere, JM. Gomez-Perez, and H. Wu. " Exploiting Linked Data and Knowledge Graphs in Large Organisations." Springer International Publishing, 2017.

[12] R. Agrawal, A. Somani, and Y. Xu. "Storage and Querying of E-Commerce Data." International Conference on Very Large Data BasesMorgan Kaufmann Publishers Inc, 2001, pp.149-158.

[13] Z. Pan, and J. Heflin. "DLDB: Extending Relational Databases to Support Semantic Web Queries." Psss1 - Practical and Scalable Semantic Systems, Proceedings of the First International Workshop on Practical and Scalable Semantic Systems, Sanibel Island, Florida, Usa, October DBLP, 2003, pp.109-113.

[14] LR. Jia, J. Liu, YU. Tong, Y. Dong, L. Zhu, B. Gao, and et al. "Construction of Traditional Chinese Medicine Knowledge Graph." Journal of Medical Informatics, 2015.

[15] R. Tong, M. Wang, H. Wang, and F. Hu. "Research on the Construction and Application of Vertical Knowledge Graphs." Knowledge Management Forum, 2016.

[16] G. Xian, R. Zhao, Y. Kou, L. Zhu, and J. Zhang. "Study and Practice on Converting and Publishing Chinese Agricultural Thesaurus as Linked Open Data.", New Technology of Library and Information Service, 2013.

[17] LX. Mao, and L. Wei. "Manual of meteorological service for staple crops" China Meteorological Press, 2015.