Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Construction and application of COVID-19 infectors activity information knowledge graph

Liming Chen a,1, Dong Liu b,1, Junkai Yang c,1, Mingyue Jiang a, Shouqiang Liu d,e, Yang Wang e

a School of Electronics and Information Engineering, Faculty of Engineering, South China Normal University, Nanhai, China
b South China Business College Guangdong University of Foreign Studies, Network and Information Technology Center, Guangzhou, China
c International Business College, South China Normal University, Nanhai, China
d School of Artificial Intelligence, Faculty of Engineering, South China Normal University, Nanhai, China
e School of Physics and Telecommunications Engineering, South China Normal University, Guangzhou, China

ABSTRACT

During COVID-19 prevention and control, people need to be aware of the outbreak situation in their area to avoid being inconvenienced by the outbreak and even becoming infected. Thus, this project constructs a knowledge graph with COVID-19 infecter activity information, by using the official flow information of the infected people from the provincial and municipal websites. This knowledge graph is the basis of the COVID-19 applications for tracing, visualization and reporting proposes. In the implementation process, we (1) collect a dataset with the information on COVID-19 cases from the prevention and control centers, (2) extract the entity elements with a Bert + BILSTM + CRF-based model, and (3) pre-process the dataset and construct a knowledge graph with manual annotation and human-based review. Finally, we use the knowledge graph to develop a web-based application to implement the question and answer, query, transmission path tracking and the "No.0" tracing infecter functions.

1. Introduction

Knowledge Graph was proposed by Google in 2012 [1], and it has been widely used in various industrial fields, for example in the Internet-oriented search fields, recommendation, Q&A systems, and semantic understanding [2–4]. There are two types of knowledge graphs: the general knowledge graphs [5,6] and the domain knowledge graphs [7]. General knowledge graphs refer to dictionary-based approaches, including WordNet [8], DBpedia [9], Zhishi.me [10] and so on. The domain knowledge graphs include financial industry graph (i.e. Ant Affair knowledge graph), enterprise knowledge graph i.e. Qichacha), medical knowledge graph (i.e. Chinese medicine knowledge platform) and so on.

Since the COVID-19 epidemic, The OpenKG organization has constructed a series of knowledge maps related to the COVID-19 [11]. Xiaomi AI Lab and Hehai University constructed the knowledge graph of COVID-19 events [12]; Xu et al. constructed the knowledge graph of COVID-19 health at Tsinghua University [13]; Zhang et al. constructed the knowledge graph of COVID-19 concept at Harbin Institute of Technology [14]; Li et al. constructed the knowledge graph of COVID-19 epidemic at IBM China Research Institute [15]; Liu et al. constructed the knowledge graph of COVID-19 materials at Wuhan University of Science and Technology [16].

To control the epidemic and provide the public with clear tracking information on the activity trajectory of confirmed COVID-19 cases, many platforms launched tracking applications related to the trajectory of COVID-19 patients in the past two years. Alipay platform launched Ali Health and Tencent WeChat launched a case patient track query function. These functions provide the basic trajectories and regional conditions that people want to know about the patient.

Due to the uniqueness and long-term nature of the COVID-19 epidemic, the virus might coexist with us for a long time. The timeliness of the information release of the epidemic is particularly critical to help people understand the epidemic situation in real-time and help them to keep away from the outbreak area. Due to the performance of knowledge graphs in interactive knowledge reasoning and discovery,
research in the direction of knowledge graphs is attracted much attention. By collecting information related to the COVID-19 epidemic, classifying and summarizing it, and obtaining the connections between events, a rich knowledge graph of COVID-19 can be constituted. The construction of the knowledge graph for patient activity information can discover the information and provide the public with a valuable reference.

In this paper, based on the patient flow information (those infected with COVID-19) on multiple provinces and cities’ websites, we use the named entity recognition model (a Bert + LSTM + CRF model) to extract the named entity elements, establish an entity-relationship-entity structure, and construct a novel knowledge graph for the flow information of the COVID-19. This knowledge graph is visualized with geographic maps for the development use of the COVID-19 applications. It provides early warning to nearby users and pushes them the location-based information about the nearby infected persons.

2. Related works

With the continuous increase in the number of COVID-19 patient cases, discovering the epidemic’s temporal and spatial transmission paths has become increasingly complex. At present, mathematical transmission models and simulation transmission models are used to discover the transmission. Specifically, the mathematical transmission models can be a set of ordinary differential equations [17], or a stochastic model with the differential equations [18]. There are approaches focusing on how infectious diseases are spread [19,20]. As the mathematical communication models are idealized models, there is a certain gap between them and reality. In order to overcome the shortcomings of the mathematical transmission models, these approaches focus on the transmission processes to establish the simulation transmission models [21,22]. However, both the mathematical transmission models and simulation models are based on the assumption of homogeneous space. They mainly study the relationship between the number of infectious cases and the change of time, focusing on mathematical statistical analysis, and ignoring the impact of spatio-temporal characteristics on the spread of infectious diseases. Much other research focuses on the transmission trend of infectious diseases from the perspective of spatio-temporal evolution [23,24], greatly improving the transmission model from the spatio-temporal dimension. However, the propagation models used for spatio-temporal information mainly consider the spatial distribution characteristics of the population. The expressions of social, semantic and temporal relationships between cases are ignored. The above three types of models mainly study the overall transmission process or trend of infectious diseases from a macro perspective rather than being able to express the transmission relationship between specific cases at the individual people. It is difficult to support the precise prevention of infectious diseases due to the inability to accurately identify the transmission paths.

In addition, knowledge graph techniques are introduced to describe concepts related to biomedical and clinical medicine information. It provides conceptual models to support the understanding and sharing of infectious disease knowledge [25]. However, the ontologies of infectious diseases constructed from the medical perspective are mainly applied to the knowledge modeling for disease treatment. It is difficult to display the population movement trend, the transmission path of cases, and the transmission and prevention processes of infectious diseases in the spatial and temporal dimensions. Therefore, it is necessary to combine the existing methods with the methods to analyze crowd activity to conduct knowledge modelling of the infectious disease transmission from the perspective of events. This is in line with the needs of infectious disease transmission analysis. At present, the medical knowledge atlas constructed with the conceptual ontology of infectious diseases has been widely used in COVID-19 prevention and treatment research. For example, OpenKG (http://openkg.cn/dataset/covid-19-concept) is collected for entities and relationships related to COVID-19 disease from the social media texts, and it further integrated the knowledge provided by Baidu Encyclopedia and Wikipedia. The COVID-19 Conceptual Knowledge Graph (http://openkg.cn/dataset/covid-19-epidemiology) was constructed, focusing on depicting the basic concepts of infectious diseases, and providing an important theoretical basis for subsequent research. However, they only focused on information from the perspective of COVID-19 epidemic medicine, lacking a conceptual system of time sequence and spatial relationship of cases. This could not adapt to the diversified description of case information in the era of big data, nor express the activity relationship between case entities.

The transmission process of infectious diseases is closely related to crowd activities. The recent research on the spatio-temporal patterns of COVID-19 transmission mainly focuses on statistical analysis, but lacks semantic analysis of transmission events. To model the cases, this paper aims to combine them to explore the dynamic evolution paths of transmission, by regarding infectious disease cases as a center,
combining the cutting-edge technology of knowledge graph, and considering spatio-temporal and semantic characteristics.

3. Methods

3.1. Process of the project

The process of the proposed work is shown in Fig. 1. Firstly, we obtain COVID-19 case flow data as experimental source data from provincial prevention and control centers, then, conduct knowledge extraction on the data to obtain preliminary knowledge representation. It includes entity extraction, relationship extraction and attribute extraction. This project uses a model of Bert + BILSTM + CRF. After the preliminary knowledge representation, the standard knowledge representation is obtained through knowledge fusion, including entity alignment and padding and attribute value alignment, and then the knowledge graph is obtained through quality assessment. We use Neo4j graph database (https://neo4j.com/) to manage the knowledge graph. It provides the support for the operations for adding or updating data. Finally, we visualize the knowledge graph and develop application interfaces based on the constructed knowledge graph.

3.2. Construction of knowledge graph

The knowledge graph construction process of this project mainly consists of eight steps: data acquisition, ontology creation, information extraction, knowledge mapping, knowledge fusion, knowledge processing, knowledge storage and knowledge update.

3.2.1. Data acquisition

In this project, the COVID-19 case activity information within a fixed time period is downloaded from the official websites of each provincial and municipal prevention and control center. Then, a MySQL database is created to store the collected data. We also develop applications to keep updating the data in real-time.

The captured data are the activity information, the activity event series, combining patient information (age, gender, symptoms, etc.), time, location, close contacts, and transportation. In order to accurately describe the transmission process of COVID-19 in detail, the six elements of the activity model (SW1H) are used to describe the patient activity (see Fig. 2). Specifically, Who refers to participants in activities, including the active and passive participants, with significant spatial and temporal characteristics. It is a necessary element to describe the patient activity, so the human-object, human-person and object-object relationships can be grasped. What refers to the current status of the participants, including suspected, confirmed, cured and dead. When refers to the time-related information, including the time of suspicion, time of diagnosis, time of illness and time of discharge. It provides information on the time dimension of the patient’s activities as well as the extent of transmission. Where refers to the location, including qualitative semantic description and quantitative coordinate range. It can also be converted into a series of trajectory datasets to show the area and spread of the patient activity. Why refers to the cause. It reflects the causal logic of the case illness. How refers to the route. It can be used as evidence to infer the type of activity and the spread of the case.

Among them, infected persons and close contacts have significant spatio-temporal characteristics and are indispensable elements for describing case activity, whereby the person-person relationship of case activity can be captured; Symptoms are descriptions of the condition of the infected person and include suspected, confirmed, cured and dead; Time refers to the time information related to the occurrence of case activity, including the time of suspicion, time of arrival at a location, time of transportation, time of confirmation, time of illness and time of discharge, which can express the information of the time dimension of case activity and thus the degree of virus transmission; Location is the place through which the infected person passes, including a qualitative semantic description and a quantitative coordinate range, which is transformed into a series of data on the path of the infected person, according to which the area of case activity and the spread of the case are represented; Event is the cause of case activity and reflects the causal logic of case diagnosis; Activity area is the pathway of case activity, which can be used as a basis for speculating the type of activity and predicting the extent of spread.

3.2.2. Creating ontology

An ontology is a collection of abstract concepts in a domain that can describe the common features of objects in the domain and the relationships between them. This section uses ontology model building, taxonomic stratification and conceptual grooming with human involvement. The ontologies are the core to construct the knowledge graph. Researchers are continuously expanding the knowledge representation in the domain of infectious diseases by studying the ontology system of infectious diseases. They define the basic concepts of infectious diseases and the relationships between concepts. However, it is inadequate to define infectious disease knowledge only according to a medical perspective. The presentation and prediction of dynamic spatio-temporal characteristics of COVID-19 cases require more information.

Ontology is an important part of knowledge graphs, which can formally represent the hierarchical relationships of concepts related to patient activities. In view of the diverse ways to describe patient activities, this paper takes the event activities of the schema layer as the core, and aims to add concepts and relationships without changing the existing ontology structure. This paper adopts the web ontology language (OWL) to represent the semantic relationships and spatio-temporal relationships of cases.

The simple event model (SEM) [26] is a generic event representation model. It does not rely on domain vocabularies and can be used to model events in different domains. The core concepts, the event, category system and attribute constraints, are used to describe events in different domains. The four concepts, time, place, object and event, are used to
describe the components of events. Attribute constraints are used to
describe attributes in the knowledge graph, which can be constrained or
extended by adding information to existing attributes. The use of the
SEM model for modelling case activity is consistent with the charac-
teristics of case activity, and a spatio-temporal event model is con-
structed to describe the conceptual model of sub-events in the activity
record.

| Semantic Relationship       | Logical Description | Instruction                                           |
|-----------------------------|---------------------|-------------------------------------------------------|
| Composed relationship       | R_{composed} (M1, E1) | Activity is composed of event E1                      |
| Cause relationship          | R_{cause} (E1, E4)  | E1 event leads to event E4 occurring                  |
| Followed relationship       | R_{followed} (E2, E3) | Event E2 follows Event E3                             |
| Concur relationship         | R_{concur} (E2, E5) | Event E2 and Event E5 concur                          |

The conceptual hierarchy needs to include the epidemiological
ontology model, the extended activity concept and the conceptual hi-
erarchy, involving the 5W1H corresponding to time, place, object,
manner, and cause elements. As the existing schema hierarchy is diffi-
cult to describe the diversity of information [27], uses an event model to
describe the conceptual model of sub-events in the activity record. The
Core Classes contain four concepts: time, participants, place and event,
describing the components of an event. The type system corresponds to
the core concepts. Property constraints are applied to the properties of
the knowledge graph. In this paper, we first expand the concepts of why
and how in relation to the event components and their corresponding
category systems, which are cg:How, cg:Why and cg:How-Type, cg:
WhyType. In addition, the temporal, spatial and spatial relationships of
events are presented in relation to the attribute constraints.
(2) Entity Correlation Design

Associative relationships in the knowledge graph schema layer include contagious relationships, social relationships, case activity event relationships, and temporal relationships. Social relations are semantic relations, including kinship and colleague relations (ⓐ in Fig. 3). Infectious relations are related to contact between cases (ⓑ in Fig. 3), related to case activity. They are rich in semantic and temporal relations.

Case activity events include gatherings (e.g. meal events), simple events (e.g. exposure events, home isolation events, medical visits), trip events (e.g. travel events, shopping events), and phenomenal events (e.g. fever events) (ⓒ in Fig. 3). In addition, the semantic relations of events contain composed, cause, followed, and concur relations. They are mainly non-hierarchical relations that can portray the contagion between cases. As shown in Table 1, the activity record M1 consists of E1, E2, E3, E4 and E5, denoted by \( R_{\text{composed}}(M1, E1), \ldots, R_{\text{composed}}(M1, E5) \), and referring to the activity-event composition relationship. A causal relationship means that the event causes other events e.g. contact event E1 leads to fever event E4, which can be denoted by \( R_{\text{cause}}(E1, E4) \). The following relationship means that an event follows another event in a certain time interval. For example, a fever event (E2) follows a travel event (E3) can be denoted by \( R_{\text{followed}}(E2, E3) \). A concurrent relationship means two events occur simultaneously within a certain time interval but they are not causally related \( (R_{\text{concur}}(E2, E5)) \), e.g. a medical event occurs while a fever event occurs.

Temporal relationships describe the temporal characteristics of infectious disease activity events through time points and time periods. It can also be used as a basis for extracting the temporal relationships of events to discover the interdependencies and mechanisms of action of events in neighboring time domains. The temporal relationships between different types of events are shown in Table 1. According to the transferability of the temporal relationships, if E1 precedes E2 and E2 precedes E3, it can be deduced that E1 precedes E3. In this case, the E1 is regarded to have a direct timing relationship with E2, and E1 has an indirect timing relationship with E3. In order to reduce the redundancy in the construction process of the temporal relationships, only direct temporal relationships between events are established.

(3) Hierarchical and event-relational representation

The activity record consists of multiple sub-events. The event elements consist of time, place, manner, participant, purpose, etc. The
participants of the event are represented using sem:Actor and sem:Object, sem:How and sem:Why describing how and why the event occurred. sem:Place and sem:Time represent the time and place information of the event, and time and place can be represented by attribute values and entities. This section represents the category information of the active event by adding concept instances without extending the existing concepts in the schema layer, and also represents the parent-child relationship of different concept instances and thus the concept hierarchy by means of hasSubType (Fig. 4), such as the <transport, hasSubType, aircraft> triplet.

We use traditional knowledge mapping to add instance relations as constraints to the concept relations within the schema layer for expressing the epidemiology knowledge mapping schema layer. Instances of the core concept (sem: Event) are connected through the instances of the model attribute constraint (sem: Role) and the two relations of subject (rdf: subject) and predicate (rdf: object), and combined with (sem:RoleType) to represent event relations, e.g. the triple <E1, cg:after, E2> can be represented by the 3 triples <cg:after, rdf: subject, E1>, <cg:after, rdf:object, E2> and <cg:after, rdf:type, sem: RoleType>. The event types are described by specific instances of the category system, e.g. the two triples <E1, sem: Event-Type, travel event> and <E1, sem: EventType, confirmed event > represent E1 and E2 as travel and confirmed events, respectively. This subsection builds on the existing ontology structure without the need to expand the schema-level relationship types, and enables a non-hierarchical relationship representation of events through existing relationship and concept types such as attribute constraints and subject-predicate relationships, which enhances the backward compatibility of the model and can better accommodate the COVID-19’s diverse description approach.

The ontology created by the knowledge graph constructed for this experiment includes infector, location and transportation entity nodes, as well as temporal relationships connecting the different nodes.

### 3.2.3. Information extraction

Information extraction is to extract knowledge such as entities, attributes, and relationships from data sources. This is the critical part of the knowledge graph construction. The quality of information extraction determines the quality of the knowledge graph. The relationship between entities and the attribute values of entities can be represented by a triple (subject, predicate, object), so information extraction can also be the triple extraction. The triple of this project is represented by (entity, attribute, attribute value), and the key technologies are entity extraction, relationship extraction and attribute extraction.

#### (1) Entity extraction

Entity extraction, also known as Named Entity Recognition (NER), refers to the automatic identification of named entities from textual datasets with the aim of creating nodes in the knowledge graph. The quality of entity extraction (accuracy and recall) greatly affect the efficiency and quality of subsequent knowledge acquisition. Therefore, it is the most fundamental and critical part of information extraction. The data of this project is unstructured textual data, which mainly includes basic patient information, patient activity records (including information of time and place, activity purpose, etc.), patient social relationship, current status and patient travel route.

This project uses a BERT + BILSTM + CRF-based named entity recognition model to perform entity extraction (Fig. 5). BERT [28] is a state-of-the-art pre-trained language model, achieving significant improvements in diverse NLP tasks [29]. The BILSTM + CRF structure consists of a BILSTM followed by a softmax layer and a CRF layer. The BILSTM accepts the embedding sequence of words in the input sentences, and the CRF layer is used to add constraints to the final prediction.

Using the Bert + BILSTM + CRF model can quickly extract the entity elements to build the knowledge graph from the case line information, and the optimized model can adapt to our data with an accuracy of over 99%.

#### (2) Relationship extraction

After the entity extraction, we obtain a series of discrete named entities (called nodes). In order to obtain the semantic information, it is also necessary to extract the association relationships (called edges) between the entities from the data. The edges in this project contain contagion relationships, case activity event relationships, and chronological relationships. The infectious relationships are a close contact of the infected person’s associated with patient activity with rich semantic and spatio-temporal relationships. Case activity events include trip events (e.g., travel events, shopping events, transportation events), and phenomenal events (e.g., fever events). In addition, the semantic relationships of events contain compositional, causal, following, and concurrent relations. They are non-hierarchical relationships, and can portray the contagion relationships among COVID-19 cases. The temporal relationships can describe the temporal characteristics of COVID-19 case activity with time points and time periods. The relationships can also be used to extract the temporal relationships of entities, and further discover the interdependencies and mechanisms of actions of entities in the neighboring time domains.

#### (3) Attribute extraction

The goal of attribute extraction is to collect attribute information of specific entities from different information sources, so, as to complete a complete outline of entity attributes, e.g., for a certain infected person, data from multiple sources (heterogeneous) can be obtained from the Internet. This project uses a sub-event conceptual model to describe the attributes of entities, where the core concept contains four concepts: time, participants, place and event to describe the constituent elements of an event.

### 3.2.4. Knowledge mapping

Knowledge mapping is to establish the mapping relationship between the structured information extracted from the underlying data and the knowledge graph ontology, while formatting editing refers to completing the knowledge mapping configuration by editing the json format code.

### 3.2.5. Knowledge fusion

Knowledge fusion is to align and merge a large number of extracted triples, identifying the same entities and merging them into one. The technical routes can be basically divided into two categories: entity attribute similarity-based and joint representation-based deep learning. Considering that the deep learning method based on joint representation relies on a large amount of annotated data, this requires a lot of labor work. This project adopts the entity attribute similarity method to complete entity alignment and knowledge fusion by defining similarity measures and combinations. This includes three tasks: entity alignment and filling, attribute value alignment and entity linking.

#### (1) Entity alignment and filling

**Entity filling.** Some entities have no attribute values due to the data sparse. We set entity filling rules to improve entity information. The column of the attribute value of the entity with no attribute value is filled with "No data for this attribute value".

**Entity alignment.** Entity alignment includes entity disambiguation and co-reference disambiguation, i.e., to determine the same entity with different names and the different entities with the same names. It also determines whether the knowledge entities relate to the same real word entity. It is to eliminate entity conflicts from multi-structured data sources to avoid unclear pointing, and other related problems. In this
The similarity function to compare two entities \((e_1, e_2)\) is defined as:

\[
sim(e_1, e_2) = (1 - \alpha)\text{sim}_{\text{structure}}(e_1, e_2) + \alpha \text{sim}_{\text{ATTR}}(e_1, e_2)
\]  

(1)

In Equation (1), \(\text{sim}_{\text{ATTR}}(e_1, e_2)\) is the attribute similarity function, \(\text{sim}_{\text{structure}}(e_1, e_2)\) is the structure similarity function, and \(\alpha\) is the adjustment parameter \((0 \leq \alpha \leq 1)\).

The attribute similarity function compares two entities according to whether their attributes are similar or not. It maps the entities to a set of characters into two n-dimensional vectors, and compares their cosine similarity (Equation (2)). This similarity reflects whether the two sets of characters are similar or not. The attribute similarity function is defined as.

\[
\text{sim}_{\text{ATTR}}(e_1, e_2) = \sum_{(\alpha_1, \alpha_2) \in \{e_1, e_2\}} \text{sim}(\alpha_1, \alpha_2)
\]  

(2)

The structural similarity function uses Jaccard correlation coefficient, which is the ratio of the intersection of the common neighbors of these two entities to the concurrent set.

\[
\text{sim}_{\text{structure}}(e_1, e_2) = \frac{|S(e_1) \cap S(e_2)|}{|S(e_1) \cup S(e_2)|}
\]  

(3)

For this project, since the acquired entity data has certain regularity, the entity alignment works are done by manual annotation and review.

(2) Property alignment

Attribute value alignment refers to the issue that values of the same attribute can have multiple expression forms. We setup unified annotation rules to reduce data redundancy and to improve the expression of knowledge expression. The following annotation rules are used.

- Multiple attribute values for the same attribute value: Separate the attribute values by a space key.
- Attribute value many-to-one problem: Attribute values with the same semantic meaning are marked out and a unified attribute value expression is used.

(3) Entity link

This experiment completes the linking of entities in the knowledge graph through entity alignment and attributes value alignment. The data is stored in json files with the form of (entity, attribute, attribute value).

3.2.6. Knowledge storage

In this project, Neo4j graph database is used to store the COVID-19 case activity elements. For the entity storage, nodes and edges are used as the description form, and storage with knowledge of two types of triples: (entity, relationship, entity) and (entity, attribute, attribute value).

3.2.7. Knowledge reasoning

After the construction of the schema layer, the prototype of a knowledge graph has been built. However, most of the relationships in the knowledge graphs are mutilated, resulting in vast missing values. Thus, this project uses knowledge inference techniques to complete knowledge discovery. Knowledge reasoning refers to the process of thinking, understanding, cognizing, analyzing and making decisions about various things from existing knowledge by using various methods to find out the implied knowledge or inferring the unknown knowledge, so that the knowledge map is gradually completed. Knowledge inference for a knowledge graph is defined as the prediction of the missing elements of triples in the knowledge graph, e.g., predicting entities and relations in an (entity-relationship-entity) triple. Entity prediction refers to the process of predicting another entity from known entities and relationships, while relationship prediction is the process of predicting the relationship, given head and tail entities. The literature [30,31] categorizes knowledge inference towards knowledge graphs into the following categories: (1) inference based on the graph structure and statistical rule mining; (2) inference based on knowledge graph representation learning; (3) inference based on neural networks; and (4)
hybrid inference.

(1) Inference based on graph structure and statistical rule mining

Among the inference methods based on the graph structure and statistical rule mining, a representative one is called Path Ranking Algorithm (PRA) proposed by Lao et al. [32]. In knowledge inference, PRA is a global algorithm based on the graph structure. It obtains the relationship paths between entities as features by random walk or traversal, calculating the feature values of samples, and adding classifiers to predict the potential relationships between entities. In the COVID-19 patient activity knowledge graph, potential transmission relationships between different patient entities can be predicted. Since then, some improved algorithms based on PRA have been proposed gradually. Gardner et al. [33] proposed an efficient and more expressive Subgraph Feature Extraction (SFE) model, which can effectively reduce the complexity of the PRA algorithm. After obtaining features from the graph structure, researchers explored the use of traditional association rule mining methods for knowledge inference. For example, Association rule Mining under Incomplete Evidence (AMIE) [34] supports the mining of closed-form rules from incomplete knowledge bases. This algorithm performs rule mining for each relation by adding hanging edges, instance edges, and closed edges in turn, and evaluates them with support and confidence as criteria.

(2) Reasoning based on knowledge graph representation learning

Methods based on knowledge graph representation learning often first feature the entities and relationships in the knowledge graph, and then use the results of the representation for knowledge inference. One of the most representative knowledge representations is the TransE approach [35], which aims to solve the problem of large-scale knowledge graph relational data processing. The method regards the relationship of each triple as a translation process from a head entity to a tail entity, and completes the representation of all the entities and relationships by learning to adjust the vector relationship between the three. The sum of vectors of the head entity and the relationship is pushed as close as possible to the vector of the tail entity. Although the principle of TransE is simple and easy to extend, it still suffers from poor modeling of complex relationships and the inability to use the information beyond the knowledge base. Later, many subsequent methods are developed upon TransE. A typical one is termed PTransE model, proposed by Zhiyuan Liu et al., which incorporates multi-step relational paths from knowledge graphs into the knowledge representation learning model.

The PTransE schematic is shown in Fig. 6. PTransE is still based on a translation assumption, while the individual relational triples are replaced by relational paths. For example, the score function defined by PTransE for the relational triples takes into account the multi-step path information between entities.

$$G(h, r, t) = E(h, r, t) + E(h, P, t)$$ (4)

Among them, $E(h, r, t)$ describes the correlation between entities and relations using the direct relation triple, such as those defined in TransE:

$$E(h, r, t) = \| h + r - t \|_{L_1/L_2}$$ (5)

In contrast, $E(h, P, t)$ is different from the PTransE model. It describes inference information at the relational level through multi-step paths. Since an entity pair $(h, t)$ may have several different relationship paths in the knowledge graph, the reliability of different relationship paths in reflecting the connection between entities varies greatly. $E(h, P, t)$ is defined as the result of the weighted average of the score function under each relationship path according to its reliability.

$$E(h, P, t) = \frac{1}{Z} \sum_{p \in P(h,t)} R(p|h, t)E(h, p, t)$$ (6)

$R(p|h, t)$ is the normalization factor. $R(p|h, t)$ and $E(h, p, t)$ weigh the reliability of the relational path and the energy of the entity pairs under the relational path, respectively. For the reliability of the relational paths, this model proposes a resource allocation algorithm that constrains the paths and measures the reliability of relational paths for this purpose.

The PTransE model embeds entities and relations in a low-latitude space by encoding relational paths. It uses path-constrained resource allocation algorithms and semantic composite algorithms to represent paths. Thus, high-performance knowledge graph complementation (entity prediction and relationship prediction) and textual relationship extraction can be achieved. However, PTransE has a disadvantage over TransE if the features of the knowledge graph are not considered by PTransE, because the entity representation of the knowledge graph provides the key information for the relationship prediction. Moreover, the hit rate of the PTransE model is only 60% in the head entity many-to-many prediction.

(3) Neural network-based inference

Neural network-based inference methods generally refer to the inference performed by using some properties of neural networks [36]. For example, predicting the missing elements in a triple, or predicting the relationship between the first and last two entities in a multi-hop path. Socher et al. [37] proposed a novel Neural Tensor Network (NTN), which represents entities as the average of word vectors within entities and has been experimentally verified to be superior to the representation of a single word vector. Lukovnikov et al. [38] proposed a neural matching model, HNM, which accomplishes the task of answering simple questions by ranking the main and predicate words. The literature [37] also showed that the knowledge inference task can be accomplished by using the storage capacity of neural networks, such as the IRN model proposed by Shen et al. [38] and the DNC model proposed by Graves et al. [39], which mainly simulate the process of human thinking and visualize the storage and reading of knowledge to accomplish the fast inference process. Neural network-based reasoning has made great progress in recent years by the virtue of its good characteristics. However, there are still problems such as insufficient explanation. Besides, it tends to focus on a single level of information in the knowledge graph, whereas cannot globally consider various influencing factors, such as semantics and paths. Last but not the least, its generalization ability needs to be improved.

(4) Mixed reasoning

In order to make up for the shortcomings of single-category reasoning methods, many scholars have begun to explore the use of multiple methods, i.e., hybrid reasoning methods. Traditional path-based methods often require a large amount of data to obtain path features. With the increasing scale of knowledge graphs, traditional methods are complex and computationally difficult, while they still have good interpretability. Neural network/representation learning-based inference methods have good computational performance but insufficient interpretability. Most hybrid inference gradually combines the two to form a variety of inference methods. Neelakantan et al. [40] proposed a relational inference model based on RNN, which was inspired by the idea of PRA method to obtain feature paths first. Then, a RNN model was used to vectorize them and finally complete the relational inference task.

3.2.8 Knowledge update

There are two ways to update the content of a knowledge graph: data-driven global updates and incremental updates. Due to the
The time, place, person and transportation elements extracted from the original data are used to build entity-relationship-entity triples. The triples construct the COVID-19 infectors activity information knowledge graph, which is shown in Fig. 7. The knowledge graph established two triples, infecter - time - place and infecter - time - transportation. Fig. 8 (a) is an infecter node relationship diagram, and Fig. 8 (b) is a location node relationship diagram. Fig. 8 (c) is a transportation node relationship diagram.

We develop the website and obtain the knowledge graph from the graph database for visualization. We set up entity query and intelligent question and answer functions on the website in Fig. 9 and Fig. 10, respectively.

5. Discussion

5.1. Advantages of the knowledge graphs and algorithmic models

This project utilized the streaming information of COVID-19 patients obtained from authoritative websites of various provinces and cities to construct a knowledge graph by scientific data preprocessing. The reliability and authenticity of the data are endorsed by the authoritative website data updated every day. We also carry out a series of cleansing and processing of the data to ensure that the most important data can be obtained to complete the precise prevention and control. The improved and optimized named entity recognition algorithm model Bert + BILSTM + CRF was used to pre-process the activity information of COVID-19 cases downloaded from provincial prevention and control centers within a limited time to extract key elements such as personal information, time, location and means of transportation of cases. Our collected data were divided into the training set, validation set and testing set with the ratio of 7:1:2. The accuracy of the model achieves 99.36%. Thus, our data extraction is more efficient and convincing.

We model the case data in the form of data organization combining spatio-temporal and semantic features, learning the rules of COVID-19 epidemic transmission, and visualizing the knowledge graph. Thus, users can more easily detect the outbreak area, know how to avoid virus infection, and learn whether their travel overlapped with those of infected people. This can meet the needs of users to avoid entering the outbreak area, or straying into the outbreak area for timely isolation and medical treatment purposes.

5.2. Features of ontology creation

In the ontology creation section, we used the six-factor model of activity (SWIH) as the basis for analysing the case activity element composition. We also adopted the Ontology Description Language

Table 7
Irrelevant characters label map.

| s | t | a | y | e | d | a | t | h | o | m | e |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Table 3
Entity extraction data results.

| Data type | Value |
|-----------|-------|
| Accuracy  | 99.36%|
| Precision | 94.14%|
| Recall    | 95.48%|
| FB1       | 94.80%|

Table 4
Person entity extraction label map.

| B-PER | I-PER | I-PER | I-PER | I-PER | I-PER | I-PER | I-PER | I-PER | I-PER |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|

Table 5
Location entity extraction label map.

| B-LOC | I-LOC | I-LOC | I-LOC | I-LOC | I-LOC | I-LOC | I-LOC | I-LOC | I-LOC |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|

Table 6
Organization entity extraction label map.

| B-ORG | I-ORG | I-ORG | I-ORG | I-ORG | I-ORG | I-ORG | I-ORG | I-ORG | I-ORG |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|

Table 2
The data results of the experiments.

| Entity type | Value   |
|-------------|---------|
| Precision   | Recall  |
| LOC         | 94.39%  | 95.76%  | 95.07%  |
| ORG         | 90.57%  | 92.71%  | 91.63%  |
| PER         | 97.97%  | 98.24%  | 98.11%  |

3.3. Knowledge graph visualization and application interfaces

By using the knowledge graph, we built a dataset based on Neo4j, and a website providing knowledge graph visualization and application interfaces. Users can use the interface to query the dataset. The website will responded with the data related to a COVID-19 case, time, and location. The application interface of the knowledge map can also be used to build intelligent Q&A systems.

4. Experimental results

The experimental results, given by the improved named entity recognition model Bert + BILSTM + CRF, are shown in Table 2. Precision, recall and FB1 values for the three types of tags (LOC, ORG and PER) in the entity extraction task are shown in Table 3. All the evaluation metrics reached high values and met the experimental data requirements.

The trained Bert + BILSTM + CRF model is applied to the experimental raw data for entity extraction (in Table 4, B-PER denotes the head of person entity, I-PER denotes the body and tail of person entity; in Table 5, B-LOC denotes the head of location entity, I-LOC denotes the body and tail of location entity; in Table 6, B-ORG denotes the head of organization entity, I-ORG denotes the body and tail of organization entity; in Table 7, O denotes irrelevant characters).

Table 7
Irrelevant characters label map.

| B-PER | I-PER | I-PER | I-PER | I-PER | I-PER | I-PER | I-PER | I-PER | I-PER | I-PER | I-PER | I-PER | I-PER | I-PER |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|

Table 8
Organization entity extraction label map.

| B-ORG | I-ORG | I-ORG | I-ORG | I-ORG | I-ORG | I-ORG | I-ORG | I-ORG | I-ORG | I-ORG | I-ORG | I-ORG | I-ORG | I-ORG | I-ORG |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|

uncertainty and real-time nature of the COVID-19 outbreak, we choose the incremental updates to add new knowledge to the existing knowledge graph, using the current new data as input. This approach is less resource intensive, whereas it still requires significant manual intervention.
(OWL) to realize the conceptual hierarchical semantic relationships and spatio-temporal relational representations of the cases. The ontology schema layer is designed in terms of three aspects: conceptual hierarchical relationships, entity association relationships and hierarchical and event relationship representations. The non-hierarchical relational representations of events through existing relationships and concept types, such as attribute constraints and subject-predicate relationships enhance the backward compatibility of the model. It can be better adapted to the diverse ways in which COVID-19 can be described. The relationships between the nodes in the knowledge graph constructed in time can visualize the behavioural paths of infectors, helping people to quickly identify risk areas, key time points and close contacts, etc., while facilitating the tracing of virus transmission paths.

5.3. Reasoning about the process of COVID-19 infector transmission

5.3.1. Relational reasoning for COVID-19 infector

We retrieve infectors by semantic information and analyze the
semantic relationships of infectors based on the case association relationship module of the prototype system, using the activity mapping of infector8 and infector9 as an example. As shown in Fig. 11, infector8 and infector9 both took subway line S9 on 7–11. The activity events of infector8 and infector9 were analysed in Fig. 11(b) and (c), respectively. It was found that they had the same itinerary on 7–11. For example, they both visited the Goodwill Shopping Centre for shopping and the Jiangning Hospital for medical treatment, which can be speculated that infector8 and infector9 may be related.

5.3.2. COVID-19 dissemination critical node analysis

The activity analysis module of the prototype system is based on the analysis of key nodes at a microscopic level, using SEM models for the occurrence of events and cases. Fig. 12 shows a complete chain of transmission. Infector3 is a patient with a long incubation period. Infector11, infector12 and Infector3 all work at Lu Kou Airport. However, infector3 was the first diagnosed, while infector11 and infector12 were diagnosed on the same day. Thus, it can be inferred that infector11 and infector12 were likely transmitted from infector3. A separate
analysis of the activities of infector11 and infector12 shows that on 7–11, infector11 and infector15 took Bus No.851 at the same time, while on 7–12, both infector12 and infector17 went to the Rui Jiang Hong Hotel for a banquet. A whole chain of propagation can therefore be deduced in Fig. 12. In this chain of transmission, infector3 was the initial node and the key node. Inadequate control of the outbreak in infector3 resulted in four confirmed cases. Many people are under medical observation.

5.3.3. COVID-19 infector temporal activity retrospective

Activity events are retrieved by the model’s event time, location and case entity. The case’s activity events are then retraced based on the prototype system’s case activity trajectory retracing module. For example, the activity profile of infector10 before a diagnosis is shown in Fig. 13. On 7–12, infector10 went to Ba Fang Restaurant for dinner; 7–13, he went to Lukou Airport for work and travelled to Lu Kou Tian Yu Fruit on 7–14. On 7–15, he went to Jiangning Hospital for medical treatment.

Due to the rich activity of the infector, one can use the case’s activity mapping and trajectory to assist in close contact identification efforts.

6. Conclusions and future work

In this experimental study, we apply the optimized named entity recognition model BERT + BILSTM + CRF to the source data to perform entity element extraction. Then, we use a hybrid knowledge inference method based on the combination of graph structure and representation learning to infer the relationships between COVID-19 case entities, the relationships between case and location entities, and the entity’s implicit properties. A knowledge graph of COVID-19 case activity information was then constructed based on the preprocessed data. However, the currently used entity extraction model and knowledge inference model still need further improvements to adapt to our dataset and to meet our requirements for entity elements and relationship elements. Besides, we will also explore how the constructed knowledge graph can be applied to text summary generation to extract and generate the needed keywords and core information of the COVID-19 epidemic from the huge amount of information.

Our direction for the future is mainly to refine the application of the project to the ground. Firstly, we will complete the application covering multiple types of users and scenarios in multiple fields, and add multiple sections such as close contacts queries, prevention and control areas and
The project was supported by Guangzhou Science and Technology Plan Project (No.201903010103), 2021 Guangdong Province Undergraduate College Teaching Quality and Teaching Reform Project Construction Project (Guangdong Higher Education [2021]29 No. 154), 2021 South China Normal University Quality Engineering Construction Project (Teaching [2021] 72No. 136), The 16th Batch of General Education Curriculum Construction Projects of South China Normal University (Teaching [2021] 74 No. 10), Guangzhou Philosophy and Social Science Planning 2022 Annual Project (2022GZYB66), 2022 South China Normal University Quality Engineering Construction project (Teaching [2022] 41 No.109 & 143), Science and Technology Plan Project of Guangdong Provincial Department of Communications (NO.2015-02-064).

The project won the competition and has been open sourced on the website: https://forgeplus.trustie.net/pomk2rwnf/COVID-19KG, competition official website https://opengcc.org/Lists/62.html.

References

[1] Google Inc, Knowledge graph. http://www.google.com/insidesearch/features/search/knowledge.html. (Accessed 29 April 2013).

[2] R. Mao, C. Lin, F. Guerin, Word embedding and WordNet based metaphor identification and interpretation, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2018.

[3] R. Mao, C. Lin, F. Guerin, Interpreting Verbal Metaphors by Paraphrasing, 2021 arXiv preprint arXiv:2104.03919.

[4] M. Ge, R. Mao, E. Cambria, Explainable metaphor identification inspired by conceptual metaphor theory, in: Proceedings of the 36th AAAI Conference on Artificial Intelligence, 2022.

[5] Q. Lin, J. Liu, Y. Pan, L. Zhang, X. Hu, J. Ma, Rule-enhanced Iterative Complementation for Knowledge Graph Reasoning, Information Sciences, 2021.

[6] Q. Lin, J. Liu, F. Xu, Y. Pan, Y. Zhu, L. Zhang, T. Zhao, Incorporating context graph with logical reasoning for inductive relation prediction, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), 2022.

[7] K. He, L. Yao, J. Zhang, Y. Li, C. Li, Construction of genealogical knowledge graphs from obituaries: multitask neural network extraction system, J. Med. Internet Res. 23 (8) (2021), e225670, https://doi.org/10.1129/e225670.

[8] G.A. Miller, WordNet: a lexical database for English, Commun. ACM 38 (1995) 31–65.

[9] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, Z. Ives, Dbpedia: A nucleus for a Web of Open Data. The Semantic Web, Springer, Berlin, Heidelberg, 2007, pp. 722–725.

[10] X. Niu, X. Sun, H. Wang, S. Rong, G. Qi, Y. Yu, Zhishi.me-weaving Chinese linking open data, in: International Semantic Web Conference, Springer, Berlin, Heidelberg, 2011.

[11] Language and Knowledge Computing Professional Committee of Chinese Information Society of China, Chinese open knowledge graph. http://openkg.cn/group/coronavirus, 2015.

[12] Language and Knowledge Computing Professional Committee of Chinese Information Society of China, Chinese open knowledge Graph.COVID-19-event. http://openkg.cn/dataset/covid-19-event, 2020.

[13] Language and Knowledge Computing Professional Committee of Chinese Information Society of China, Chinese open knowledge Graph.COVID-19-health. http://openkg.cn/dataset/covid-19-health, 2020.

[14] Language and Knowledge Computing Professional Committee of Chinese Information Society of China, Chinese open knowledge Graph.COVID-19-concept. http://openkg.cn/dataset/covid-19-concept, 2020.

[15] Language and Knowledge Computing Professional Committee of Chinese Information Society of China, Chinese open knowledge Graph.COVID-19-epidemiology. http://openkg.cn/dataset/covid-19-epidemiology, 2020.

[16] Language and Knowledge Computing Professional Committee of Chinese Information Society of China, Chinese open knowledge Graph.COVID-19-goods. http://openkg.cn/dataset/covid-19-goods, 2020.

[17] A.L. Lloyd, Realistic distributions of infectious periods in epidemic models: changing patterns of persistence and dynamics, J. Theor. Popul. Biol. 60 (18) (2001) 59–71.

[18] X.J. Xu, W.X. Wang, T. Zhou, et al., Geographical effects on epidemic spreading in scale-free networks [J]. Int. J. Mod. Phys. C 17 (12) (2006) 815–822. 822.

[19] M. Lipsitch, T. Cohen, B. Cooper, et al., Transmission dynamics and control of severe acute respiratory syndrome [J]. Science 300 (5 627) (2003) 1 966–1 970.

[20] S. Riley, C. Fraser, C.A. Donnelly, al. et, Transmission dynamics of the etiological agent of SARS in Hong Kong: Impact of public health interventions[J]. Science 300 (5 627) (2003) 1 961–3 966.

[21] W.D. Leslie, R.C. Brunham, The dynamics of HIV spread:A computer simulation model[J]. Comput. Biomed. Res. 23 (4) (1990) 380–401.

[22] M.A. Fuentes, M.N. Kuperman, Cellular automata and epidemiological models with spatial dependence[J]. Physica A 267 (3) (1999) 471–486.

[23] S. Eubank, Scalable efficient epidemiological simulation[ C], in: The 2002 ACM Symposium on Applied Computing, Spanish, Madrid, 2002.

[24] L. Gong, X.N. Liu, X.Y. Zou, Spread infectious disease risk assessment based on the spatial–temporal trajectory data analysis [J], Acta Geologica Sinica Daetica et Cartographica Sinica 44 (8) (2015) 6–12.

[25] M.X. Feng, Z.X. Fang, X.B. Lu, et al., Traffic analysis zone -based epidemic estimation approach of COVID -19 based on mobile phone data: an example of wuhan [J], Geomatics Inf. Sci. Wuhan Univ. 45 (5) (2020) 651–657, 681.

[26] W.R.V. Hage, V. Malaise, R. Segers, et al., Design and use of the simple event model (SEM)[J]. Social. Sci. Electron. Publ. 9 (2) (2011) 128–136.

[27] J.N. Liu, H.Y. Liu, X.H. Chen, et al., Terrorism event model by knowledge graph [J], Geomatics Inf. Sci. Wuhan Univ. (2020), https://doi.org/10.13203/j.whugis201904.versity, 2020.

[28] J. Devlin, M.W. Chang, K. Lee, et al., Bert: Pre-training of Deep Bidirectional Transformers for Language understanding[J]. 2018 arXiv preprint arXiv:1810.04805.

[29] R. Mao, X. Li, Bridging towers of multi-task learning with a gating mechanism for aspect-based sentiment analysis and sequential metaphor identification, in: Proceedings of the 25th AAAI Conference on Artificial Intelligence, 2021.

[30] Z.G. Ma, B.Y. Ni, K.H. Yu, Recent advances, key techniques and future challenges of knowledge graph, Chin. J. Eng. 42 (10) (2020) 1254–1266.

[31] S.P. Guan, X.L. Jin, Y.T. Jia, et al., Knowledge reasoning over knowledge graph: a survey, J. Softw. 29 (10) (2018) 2966–2994, https://doi.org/10.13203/j.cnki.jos.005551.

[32] N. Lao, W.W. Cohen, Relational retrieval using a combination of path-constrained random walks, Mach. Learn. 81 (1) (2011) 53–67, https://doi.org/10.1007/s10994-010-5205-8.

[33] M. Gardner, T. Mitchell, Efficient and expressive knowledge base completion using subgraph feature extraction, in: Proceedings of 2015 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Lisbon, 2015, pp. 1488–1498, https://doi.org/10.18653/v1/D15-1196.

[34] L.A. Galárraga, C. Tefiouni, K. Hose, et al., AMIE: association rule mining under incomplete evidence in ontological knowledge bases, in: Proceedings of the 22nd International Conference on World Wide Web, ACM, Rio de Janeiro, 2013, pp. 413–422.

[35] A. Bordes, N. Usunier, A. Garcia-Duran, et al., Translating embeddings for modeling multi-relational data, in: Proceedings of the 26th International Conference on Neural Information Processing Systems. Lake Tahoe: ACM, 2013, pp. 2787–2795.

[36] D. Lukovnikov, A. Fischer, J. Lehmann, et al., Neural network based question answering over knowledge graphs on word and character level, in: Proceedings of the 26th International Conference on World Wide Web, ACM, Perth, 2017, pp. 1211–1220.

[37] R. Socher, D.Q. Chen, C.D. Manning, et al., Modeling large-scale structured relationships with shared memory for knowledge base completion, in: Proceedings of the 26th International Conference on Neural Information Processing Systems, 2013, pp. 926–934. Lake Tahoe: ACM.

[38] Y.L. Shen, P.S. Huang, M.W. Chang, et al., Inference over knowledge graphs on word and character level, in: Proceedings of the 26th International Conference on Neural Information Processing Systems, 2013, pp. 57–67. Lake Tahoe: ACM.

[39] A. Neelakantan, B. Roth, A. McCallum, Compositional vector space models for knowledge base completion, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics, 2015, pp. 156–166.