Knowledge Graph for China's Genealogy[1]

Xindong Wu, *Fellow, IEEE*, Tingting Jiang, Yi Zhu, and Chenyang Bu

Abstract—Genealogical knowledge graphs depict the relationships of family networks and the development of family histories. They can help researchers to analyze and understand genealogical data, search for genealogical descendant paths, and explore the origins of a family. However, the heterogenous, autonomous, complex, and evolving natures of genealogical data bring challenges to the development of contemporary genealogical knowledge graph models. Applying existing methods to genealogical data may be improper because general knowledge graph models lack in-depth domain knowledge. In this paper, we propose a genealogical knowledge graph model named Huapu-KG that combines HAO intelligence (human intelligence + artificial intelligence + organizational intelligence) to implement the construction and applications of genealogical knowledge graphs. Furthermore, challenges in constructing genealogical knowledge graphs are demonstrated, and experiments conducted on real-world genealogical datasets verify the feasibility and effectiveness of our proposed model.

Index Terms—Genealogy, knowledge graph, HAO intelligence

1 Introduction

Knowledge graphs originate from the semantic web, are similar to semantic networks in essence, and have drawn extensive attention from academia and industry in recent years. Existing knowledge graphs for the general domain are mostly built from internet encyclopedias, such as Wikipedia and Baidu Encyclopedia, which prove their advantages in supporting general applications, i.e., from information extraction [13] and entity alignment [24], to question answering [3], [4]. However, many domain-specific applications, including, genealogy, are not well supported by these generic knowledge graphs, because in-depth domain knowledge is required.

A genealogical knowledge graph system aims to collect and integrate genealogical information, describe genealogical knowledge by establishing relationship networks, and visualize family evolutions, which can help users understand the origin and history of their family. However, few related studies have been conducted on genealogical data, because of the following problems. 1) A large quantity of genealogical data is recorded using traditional paper-based resources. These data must be effectively integrated with other data when building a knowledge graph model; for instance, they may need to be converted into electronic form, which is a costly procedure [1]. 2) Genealogical data are a typical example of the HACE characteristics [31] - heterogeneous, autonomous, complex, and evolving. Genealogical files can exist in horizontal, vertical, and even tree structure forms [4]. A unified algorithm or program may struggle to process these data. Moreover, some genealogical data may be missing, further increasing the difficulty of processing genealogical datasets. 3) Genealogical data are highly specialized. Chinese genealogical culture has a long history, and different dynasties have established different archiving mechanisms for the preservation and development of genealogical data. Novel words and definitions are constantly emerging, making genealogical data less readable.

To address the above problems, we propose a genealogical knowledge graph model named Huapu-KG (Huapu Knowledge Graph). With the support of HAO intelligence (human intelligence (HI) + artificial intelligence (AI) + organizational intelligence (OI)) [26], [27], heterogeneous genealogical data are collected, integrated, cleaned, and represented to form a coherent knowledge graph. Specifically, 1) AI technologies including optical character recognition and data governance are used to extract, clean, standardize, and integrate genealogical data. 2) HI, i.e., domain experts, with the assistance of OI - the capability of a genealogical organization to create and use genealogical knowledge graphs, constructs domain knowledge bases and provides domain specifications and achieving standards for

1. A shorter version of this paper won the Best Paper Award at IEEE ICKG 2020 (the 11th IEEE International Conference on Knowledge Graph, ickg 2020.bigke.org).

- *Xindong Wu is with the Key Laboratory of Knowledge Engineering with Big Data, Hefei University of Technology, Ministry of Education, Hefei, Anhui 230009, China, and with the Research Institute of Big Knowledge, Hefei University of Technology, Hefei, Anhui 230009, China, and also with the Mininglamp Academy of Sciences, Mininglamp Technology, Beijing 100102, China. E-mail: xwu@hfut.edu.cn.*
- *Tingting Jiang and Chenyang Bu are with the Key Laboratory of Knowledge Engineering with Big Data, Hefei University of Technology, Ministry of Education, Hefei, Anhui 230009, China, and with the Research Institute of Big Knowledge, Hefei University of Technology, Hefei, Anhui 230009, China, and also with the School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, Anhui 230009, China. E-mail: jiangtt@mail.hfut.edu.cn, chenyangbu@hfut.edu.cn.*
- *Yi Zhu is with the School of Information Engineering, Yangzhou University, Yangzhou, Jiangsu 225009, China, and with the Key Laboratory of Knowledge Engineering with Big Data, Hefei University of Technology, Ministry of Education, Hefei, Anhui 230009, China, and also with the Research Institute of Big Knowledge, Hefei University of Technology, Hefei, Anhui 230009, China. E-mail: z8d1177@126.com.*

Manuscript received 31 October 2020; revised 19 February 2021; accepted 8 April 2021. Date of publication 20 April 2021; date of current version 7 December 2022.
(Corresponding author: Xindong Wu.)
Recommended for acceptance by G. Chen.
Digital Object Identifier no. 10.1109/TKDE.2021.3073745

genealogical knowledge. 3) HI, AI, and OI collaborate and interact with each other to construct genealogical knowledge graphs and applications.

This manuscript is an extended version of our previous conference paper [28]. The differences between this study and the previous study are as follows:

- We propose a genealogical knowledge graph model called Huapu-KG to implement the construction of a genealogical knowledge graph with HAO intelligence.
- We verify the feasibility of the proposed model by implementing a real-world case study in the China's Genealogy website (https://www.zhonghuapu.com/).
- We examine the demonstration scenarios of genealogical knowledge graphs, to illustrate the practicality of the system.

The remainder of this paper is structured as follows. Section 2 reviews the background research. Section 3 introduces the overview of Huapu-KG. Section 4 presents the design details of the Huapu-KG model. Section 5 evaluates and analyzes our model. Section 6 discusses its real-world application. Section 7 summarizes the paper.

2 Background

In this section, we review the literature pertaining to knowledge graphs, knowledge of kinship, and HAO intelligence respectively.

2.1 Knowledge Graph

Google introduced the concept of a knowledge graph in 2012; it is a structured, semantic knowledge network that reflects the complex relationships between entities in the real world [21], [25]. Recently, knowledge graphs have been widely applied and developed into crucial technical support for semantic-searching, question-answering, and recommending systems.

Most early knowledge graphs were manually constructed by professionals. These knowledge graphs—for instance, Wordnet [20] and Hownet [10]—are small in scale but of high quality. As the scale of data increased, researchers began to construct knowledge graphs automatically—for example, YAGO [23] and DBpedia [2]. However, such knowledge graphs are often incomplete or noisy owing to inaccuracies in the extraction process and the multi-source heterogeneity of the data. There have been several attempts to construct knowledge graphs in China. For example, Baidu–Zhixin schemas[2], developed by the Baidu company, visualizes knowledge from multiple perspectives, helping users to obtain information more accurately and conveniently. The knowledge factory laboratory of Fudan University has developed and maintained a lightweight knowledge graph of a Chinese general encyclopedia called CN-DBpedia [32].

Meanwhile, knowledge graph models are in their infancy in the field of genealogy. In different countries, owing to cultural differences, kinship terms are not as detailed as those in China. Therefore, few studies have been conducted in the field of Chinese genealogical knowledge graphs. The "ancestry.com" collected multiple databases, including census data, military service records, criminal records, and church records in the United States, Britain, and elsewhere. Users can chart their family's migration history visually on a map. However, Ancestry is a for-profit company with some genealogy-related services, and owing to the above cultural differences, their website cannot be directly applied to Chinese genealogy. In China, many studies have been conducted on the history or families of well-known figures. For example, scholars have established an expert question-and-answer system concerning the complex relationships between characters in the Chinese novel "A Dream of the Red Chamber" by Cao Xueqin. However, this knowledge is not universal because it targets a specific family. Some enterprises have also developed genealogy retrieval websites, such as Root Search Network[3] and China Genealogy Network[4]. However, the complete process of genealogical knowledge graph construction and application is not given.

2.2 Knowledge of Kinship

"Genealogy" refers to documents and books recording the reproduction and evolution of a family; they have significant historical and cultural value. Kinship is key in genealogy. Typically, kinship occurs through marriage and childbirth and includes spouses, blood relations, and relations by marriage [11].

Spouse. "Spouse" refers to a couple in marital relations, and it is the basis of other kinship relations.

Blood Relation. "Blood relation" refers to a relative related by blood. Blood relations include lineal and collateral relatives. A lineal relative by blood is a relative who is directly related by blood; in general, this relation includes two categories. On the one hand, it includes parents, grandparents, maternal grandparents, and the great-grandparents of higher generations. On the other hand, it involves children, grandchildren, and lower-level great-grandchildren. Collateral relatives by blood are relatives who have an indirect blood relation.

Relation by Marriage. Relations by marriage arise from marriages. They can be divided into three main categories: (1) the spouse of a blood relative, such as a sister-in-law; (2) a spouse's blood relatives, such as a father-in-law; and (3) the spouse of a spouse's blood relative, such as the wife of a wife's brother. It is important to note that relations by marriage change with marital status.

2.3 HAO Intelligence

Based on [26], Wu *et al.* [27] proposed the "HAO intelligence" model for big data governance. The model starts from big data and provides data governance support based on HAO intelligence, which is coordinated by human intelligence, artificial intelligence, and organizational intelligence. The HAO model supports multi-type and multi-source data extraction and transmission, as well as different types of extraction and aggregation task configurations.

2. http://yingxiao.baidu.com/product/site/zhixin/

3. http://www.xungen.so/

4. http://www.fankhome.com/

Meanwhile, the model can also facilitate the implementation of data specification rules and the automatic analysis of heterogeneous data based upon data elements. The HAO model consists of three core modules.

Data Access Module. First, data acquisition tools (based on cloud computing and distributed storage) are employed to extract, integrate, process, transform, and load structured, semi-structured, and unstructured resources. The data extraction of this module incorporates three methods: panoramic, incremental, and real-time extraction. The extracted data are imported into the data sink. For other database systems, the data can be imported into the data sink via a data-exchange platform.

Data Governance Module. This module performs data cleaning and data specification on the data in the aggregation library, as well as subject division, data association (when necessary), and data integration. After this step, the data are fed into the data-sharing center.

Data Service Module. The data service module provides multi-channel and multi-dimensional data services to different users, based on the knowledge graph. On the one hand, it provides users with model management, intelligent discovery, model exploration, data subscription, and other services. On the other hand, it also provides mining analysis and expert modeling for professionals.

3 Overview of Genealogical Knowledge Graph

Methods for constructing knowledge graphs include logical modeling, hidden space, human-computer interaction, and ontology model. Human-computer interaction contains three modes. The first mode involves a computer program that questions human experts. Specifically, when a question is answered completely, the concepts and relationships about the question are modeled to generate a knowledge graph. The second mode is structured interactive knowledge traction, which refers to the description and traction of knowledge in describing the entity form the representation method of knowledge object. The third mode is to try to describe the concepts and relationships involved in human intelligence, artificial intelligence, and organizational intelligence. Moreover, pre-defining entity and relation by utilizing the synergistic effect of human intelligence, artificial intelligence, and organizational intelligence can greatly improve the quality and efficiency of genealogical knowledge graph because of the complex kinship semantics. Several generic entities in the family tree and relations between them are shown in Fig. 1. The entities include genealogy (which consists of public genealogy, joint genealogy, and private genealogy), person, and user to be specific. In addition, the attributes of persons, i.e., birthplace, should be allowed to extend for practical designs. The relations include the matching relationships between users and persons, the subordinate relationships between persons and genealogies, and the kinships and social relationships between persons. These relations between entities are the key components of knowledge graph analysis and application. For example, analyzing and modeling the relation between husband and wife would be helpful to the study of family marriage and family name association.

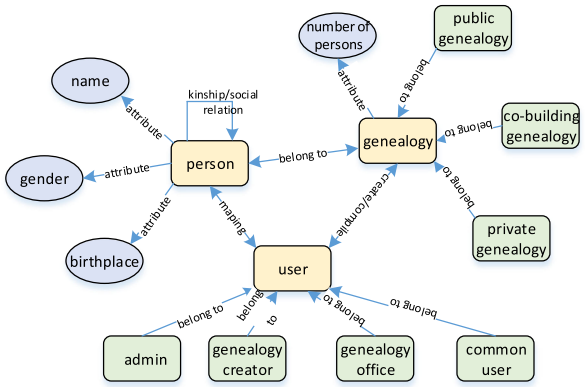


Fig. 1. Entity and relation model of the Huapu-KG (some entities and relations are not displayed in the figure).

The overall architecture of the proposed genealogical knowledge graph model (as shown in Fig. 2) mainly includes: extracting, denoising, fusing, representing, querying, applying genealogical knowledge from heterogeneous data sources based on the HAO intelligence technology.

HAO Intelligence. HAO intelligence contains human intelligence, artificial intelligence, and organizational intelligence. It runs through the whole process of the construction of a genealogical knowledge graph and provides the theoretical basis and technical support.

Genealogy Data Sources. All types of data sources related to the field of genealogy, including paper-based resources, pictures, and other structured and semi-structured data sources. It provides data support for the construction of a genealogical knowledge graph.

Genealogical Knowledge Acquisition. Genealogical knowledge acquisition is the first step in the construction of a knowledge graph. Specifically, it is a technique for automating or semi-automating the extraction of structured information such as entities, relations, and attributes from genealogical data sources.

Genealogical Knowledge Denoising. Genealogical knowledge denoising refers to the detection and completion of error and missing information in the genealogical knowledge graph, which is the first step to improve the quality of knowledge.

Genealogical Knowledge Fusion. Genealogical knowledge fusion refers to the merging of data from different sources. The goal of knowledge fusion is to improve the hierarchy and logic of genealogical knowledge. It is a key step to improve the quality of knowledge.

Genealogical Knowledge Representation. Genealogical knowledge representation interprets our cognition and expression of genealogical knowledge to achieve the purpose of knowledge sharing. Knowledge representation is the core of knowledge-based artificial intelligence applications.

Genealogical Knowledge Query. Genealogical knowledge query provides interfaces to access data for us based on formal query language. Because knowledge graphs are stored through a database system. In addition, graph query technology can be used to search specific subgraphs, namely subgraph matching.

Genealogical Knowledge Application. Genealogical knowledge graph provides the core technology for application, i.e., surname association, genealogical data visualization,

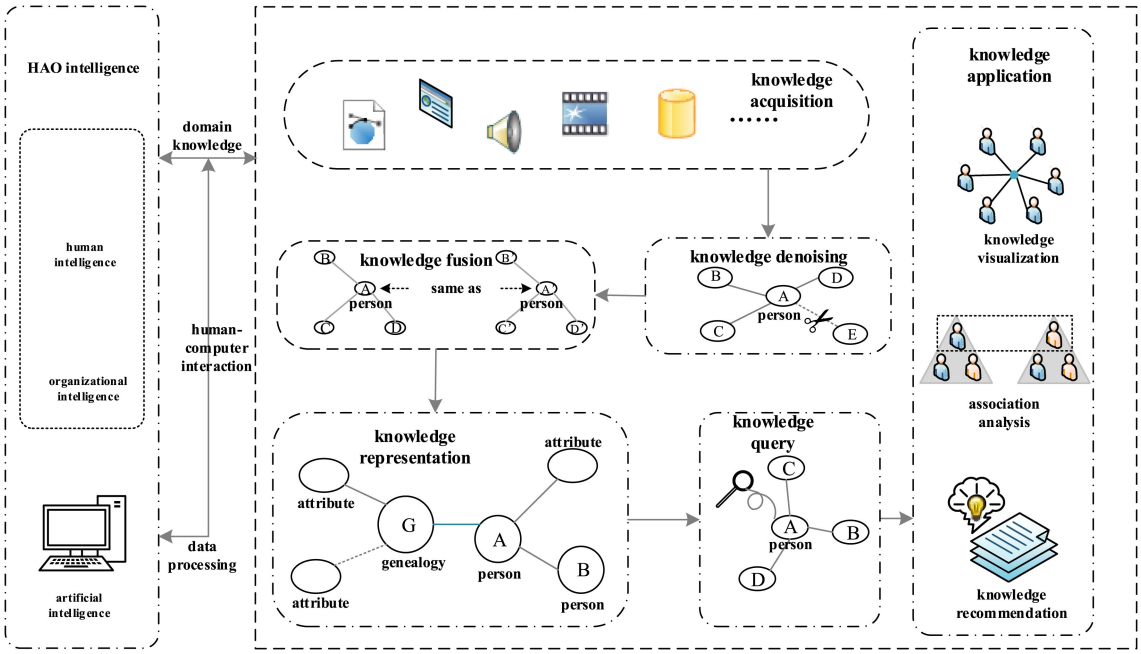


Fig. 2. Overall construction and application framework of Huapu-KG.

and social system. Genealogical knowledge application is the top level of the life cycle of the genealogical knowledge graph.

4 Genealogical Knowledge Graph Construction

Next, we will introduce the key components of the genealogical knowledge graph construction in detail.

4.1 Genealogical Knowledge Acquisition

4.1.1 Challenges of Genealogical Knowledge Acquisition

The content of genealogical data from a single source is usually short and contains insufficient semantic information [30]. In other words, the genealogical data are usually sparse and there is a lack of labeled information. Therefore, the acquisition of genealogical data is more difficult than that of general data, and existing supervised learning methods based on deep learning models may not be applicable [14], [18]. For example, in the genealogy shown in Table 1, the first line "吴德 钱文欣 二子:丰、林 (WuDe Qian Wenxin son: Feng, Lin)" describes the wife and son information of the person "WeDe". We can observe that the semantic information of the text is incomplete. It is difficult to achieve satisfactory data extraction accuracy through a supervised deep learning model. Therefore, to obtain genealogical data, unsupervised and self-supervised methods should be utilized priority, i.e., rule-based methods [33]. However, establishing rule bases for each genealogy is required, which would consume considerable expenditure of resources and is not conducive to the development of the practical application.

TABLE 1
The Wu Family Tree

十三世	
吴德 钱文欣 二子:丰、林	
吴智 赵氏 三子:松、恕、伦	
吴强 李氏 一子:恺	
Thirteenth	
WuDe Qian Wenxin son: Feng, Lin	
Wuzhi Zhao son: Song, Shu, Lun	
WuQiang Li son: Kai	

4.1.2 Our Proposed Method

A general genealogical information extraction method is introduced in our model. The synergy of human intelligence (HI) and artificial intelligence (AI) is used to extract information from genealogical data under the industry standards provided by organizational intelligence (OI). More specifically, first, the person's name is obtained from the pedigree chart with human intelligence (HI), which provides supervision for the extraction algorithm and can improve data extraction accuracy. Second, a global knowledge base is constructed by domain experts (HI + OI). When processing a specific genealogy, the genealogy information is initially extracted according to this constructed global knowledge base. Then, the unsatisfactory extraction results would be marked by humans (HI) and modified automatically by computers (AI), which effectively avoids rewriting rules for genealogies with similar structures. We can get a universal and reliable knowledge base in the field of genealogy. The detailed process of genealogical knowledge acquisition is shown as follows.

1) Data Source. The sources of genealogy data mainly including:

Traditional paper-Based Resources. Because of the underdevelopment of network technology, ancient genealogies were mainly kept and passed on in the form of paper materials.

Genealogies in the Form of Pictures and Photos. Along with the advancement of digital technology, some genealogical data have been converted into photos or pictures for efficient storage.

Genealogical Files on the Internet. Electronic genealogy compiling has become popular recently. The electronic version of a family tree can be easily compiled and published on the Internet and various genealogy websites can be utilized to edit family trees conveniently. Therefore, there is abundant public genealogical data on the Internet.

2) Data Collection. Four methods for genealogical data collection are proposed with a data privacy security protocol.

Direct Data Input. The predefined templates are utilized for entering genealogical data. Specifically, three styles of templates are defined including the table formats, the text formats, and the user interface. This method is intuitive but labor-intensive and time-consuming.

Input via Typing. The keyboard operators organize the paper-based genealogy resources into a program readable format at first. Then, these data with a specific format are processed by the machine.

Optical Character Recognition (OCR). The OCR technology is used to scan the paper-based genealogy resources to generate electronic genealogies. In certain cases, some steps for image preprocessing, such as binarization, are also required to improve the identification accuracy. Moreover, the identification accuracy could reach 95 percent with some mature OCR tools.

Web Crawler. To collect data from online encyclopedia and genealogy websites, the web crawler technology is adopted.

3) Data Extraction. Relying on computers to process and analyze genealogical data directly is less efficient. The reason is that genealogical data contain a large domain vocabulary and special grammar. It is observed that genealogical data has certain structures and regularity in the writing and layout, but the writing methods and layout structures of different genealogies are not completely the same. Therefore, a general genealogical information extraction method is proposed based on the HAO intelligence, which can avoid defining rules for different genealogies repeatedly. More specifically, human intelligence, i.e., domain expert, summarizes the language patterns of genealogical data through the understanding and analysis of genealogical data. To evaluate the quality of language patterns, we adopt sampling frequency and sampling accuracy to measure language patterns.

Sampling Frequency. Sampling frequency refers to the number of occurrences of language patterns in several randomly selected genealogical text fragments. A higher sampling frequency indicates a better language pattern.

Sampling Accuracy. Sampling accuracy refers to the proportion of the correct results in several randomly selected genealogical text fragments. Similarly, higher sampling accuracy indicates a better language pattern.

Then a global knowledge base is built based on defined language patterns to provide computers with semantic knowledge in the field of genealogy. Next, rough knowledge is extracted by the computer and fed back to the humans. If humans are satisfied with the extraction results, the results are stored. Otherwise, humans modify or add the language patterns for extracting again. Examples of language patterns are described in Table 2. Where N, SN, FN, WN denotes a male person, the son of the person, the father of the person, and the wife of the person, respectively.

TABLE 2
The Language Patterns

	language patterns
pattern 1	N: eldest son SN, second son SN, third son SN, fourth son SN, . . .
pattern 2	N, FN' son, courtesy name **, pseudonym **, born in **, died in **, wife WN, son SN, . . .
pattern 3	N WN, WN, . . . son SN, . . .

4.2 Genealogical Knowledge Denoising

4.2.1 Challenges of Genealogical Knowledge Denoising

The authenticity of genealogical data is difficult to distinguish because of the noises generated in the process of editing genealogy. To ensure the integrity and accuracy of genealogical data, regular improvements and updates are required. However, missing or corrupted resources due to improper keeping and staff members who are not familiar with the business both create noises. For example, the same person may have different names in different versions of a family tree. We cannot recognize whether the difference is caused by the change of person's name or a clerical error in writing based on the content of the family tree. Moreover, the dynamic nature of genealogical data, such as divorce, adoption, and promotion, further makes it difficult to distinguish between correct and false genealogical data. For example, the same person may have different spouses in different versions of a family tree. We also cannot distinguish whether the difference is caused by the changes in a person's marital status or noises.

4.2.2 Our Proposed Method

It is difficult to detect and correct the noises completely in a large-scale knowledge graph. Therefore, we target two types of noise data including false relations and attributes. The reasons are analyzed as follows: 1) error relations would affect the construction of kinship networks and personal path reasoning. For example, the error relations between persons and themselves would result in the endless loop of personal path reasoning. 2) The missing attributes, i.e., the "name" attribute is missed, which may make it impossible to display the complete family tree. Because the nodes of the family tree are identified by names. In addition, dealing with the two types of noises and ignoring other minor noises, i.e., personal experience descriptions, can save manpower and material resources while ensuring effective knowledge services. Specifically, the prior knowledge rule bases are defined based on human intelligence and organizational intelligence jointly and artificial intelligence techniques, i.e., subgraph matching, are used for detecting noises. Then the detection results would be fed back to human intelligence for correction, eventually, the correct data would be added to the knowledge graph. The

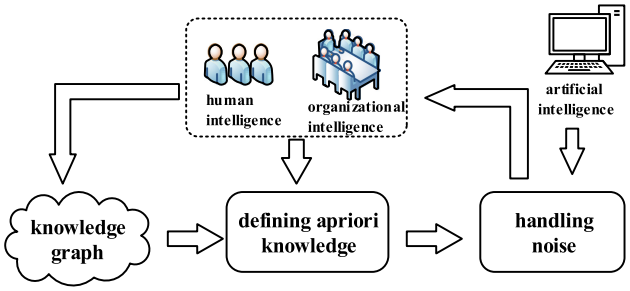


Fig. 3. The overall framework of genealogical knowledge denoising.

TABLE 3
The Prior Knowledge for Relation Definition

rule 1	If A and B are husband and wife; A is a son (or daughter) of C, Then B can't be a daughter (or son) of C.
rule 2	If A is the husband of B, Then B is the wife of A.
rule 3	If A and B are husband and wife; A is the father of C, Then B is the mother of C.
rule 4	There can't be any kinship between A and A.
rule 5	If A is the father of B, Then A's gender is male.

overall framework of genealogical knowledge denoising is shown in Fig. 3.

1) Defining Prior Knowledge Rule Base. Prior knowledge refers to the knowledge of objective existence which is independent of experience, i.e., the gender of the father is male. The prior knowledge rules about the kinships between persons in the genealogical knowledge graph are shown in Table 3.

Rule 1 describes the relations between three persons, i.e., when two persons have a definite relation with the same person, they can't establish some relationship with each other. Rule 2 illustrates the reflexivity of the relations and Rule 3 describes the transitivity of the relations. Rule 4 states that persons can't establish a kinship with themselves. Rule 5 explains the association between relations and attributes, that is, some relations constrain the value of some attributes.

2) Dealing With Errant Relations. Subgraph matching techniques, i.e., the exact subgraph matching algorithm VF3 [6], are utilized to detect the errant relations in the genealogical knowledge graph. The algorithm searches for subgraphs (i.e., errant relations) in the data graph (i.e., genealogical knowledge graph) that are exactly consistent with the schema graph (i.e., the errant relation patterns defined in Fig. 4). Therefore, the key step for improving the performance of this algorithm is defining and constructing the accurate schema graph. The four types of errant relation schemas are introduced in this paper, out of which nodes represent persons and edges denote relations. Next, we detail the four errant relation schemas.

Relation Error. Relation error includes schema 1, schema 2, schema 3, and schema 6. Specifically, schema 1 denotes that if A and B are husband and wife, and C is the father of A, then the "father-son" relation between B and C is wrong. Schema 2 states that if A and B are husband and wife, and C is the mother of A, then the "mother-son" relation between B and C is wrong. Schema 3 describes that if A is the husband of B, then B can only be the wife of A. Schema 6 states that if A is the father of B, then it is wrong that the gender of A is female. In addition, schema 1, schema 2, and schema 3 reflect rule 1 and rule 2. Schema 6 reflects rule 5.

Relation Missing. If A and B are husband and wife, and C is the son of A, then the "father-son" relation should be established between B and C. This schema reflects rule 3.

Relation Self-Loop. Relation self-loop results from the kinship established between A and A. This schema reflects rule 4.

3) Dealing With Errant Attributes. We hope to correct errant attributes at little risk when the values of these attributes don't meet expectations. Therefore, we only do corrections for attributes whose values can be normalized. To be specific, the entity set whose attribute value does not meet the normalized format should be generated first. Next, we set these errant attribute values to empty and the mechanism of "attribute padding" is adopted to padding these errant attribute values. For example, the gender of the entity "Wu Deyi" is "Beijing", which does not satisfy the requirements of normalization. It is believed that the value of "Beijing" is incorrect and the value would be set to empty. Then the following padding mechanism is executed for the "gender".

If there are edges with the label " wife", "daughter" or "mother", the gender is "female".

If there are edges with the label " husband", "son", and "father", the gender is "male".

4.3 Genealogical Knowledge Fusion

4.3.1 Challenges of Genealogical Knowledge Fusion

The same person can be called different names in different genealogies. For example, "WuDe" and "Wu Xiangfu" both refer to the same person, i.e., WuDe. Moreover, the same person can be described differently in different

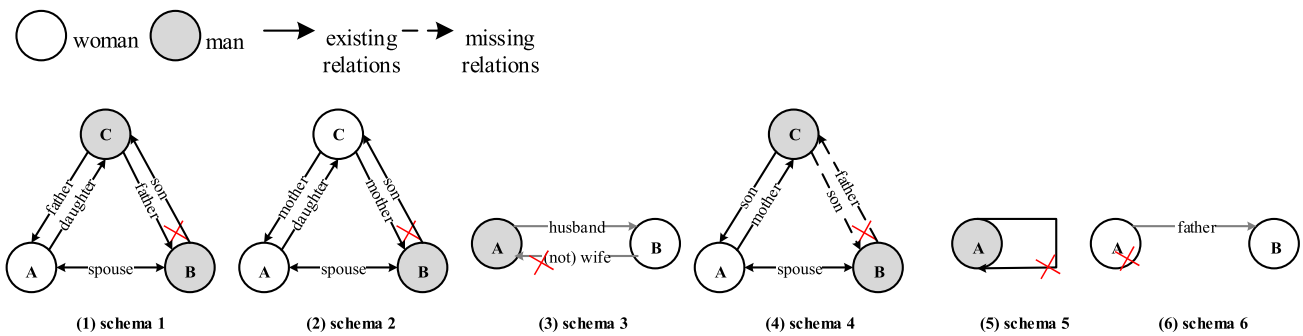


Fig. 4. Errant relation schemas.

TABLE 4
The Rules for Generating Candidate Persons

rule 1	The last name and first name of the two persons are exactly the same. For example, "San Zhan" and "San Zhan" refer to the same person.
rule 2	The last name of two persons is the same, but the first name is not exactly the same. It is observed that the name of a person in a family tree usually consists of "surname + generation + first name", but sometimes the name of a person is only "surname + first name". For example, "Wu Zizhong" and "Wu Zhong" refer to the same person.
rule 3	Two persons have the same surname, and one person's first name is exactly the same as the other's alias. For example, the tang dynasty poet Li Bai has the alias "Taibai", so "Li Taibai" also refers to "Li Bai".
rule 4	The two persons have different last names and the same first names. For example, "Luo Yi" is renamed "Li Yi" because of royal bestowal.
rule 5	Two nodes represent the same person, then its offspring will most likely represent the same person.

genealogies. Thus, adopting existing algorithms to align genealogical person could lead to the following issues. Existing methods for entity alignment can be basically divided into two categories: symbol-based methods and embedding-based methods. The symbol-based methods calculate the similarities between entities based on literal information, such as string similarity. These methods may not be able to deal with the situation of literal heterogeneity directly, namely, the inconsistency between the appellation of " WuDe" and " Wu Xiangfu". The embedding-based methods learning the similarities between entity vectors according to the structure information and attribute information between entities. However, the description and structure of the same person in different genealogies may be various, which hinders the embedding-based models to capture complete and correct semantic information between entities in different genealogies. As a result, applying embedding-based methods to genealogical personal fusion directly may be inefficient.

4.3.2 Our Proposed Method

Human intelligence technologies are utilized to obtain the rule sets of candidate persons based on the analysis of a large of genealogical data, which can deal with the case of literally heterogeneous effectively. The specific scenes for candidate persons are detailed in Table 4. In addition, to capture the semantic information of persons more precisely, the weighted sums of similarity between relations and similarity between attributes are regarded as the total similarity score for personal fusing. In this way, the same persons can be well-identified through specific kinships, even if the attribute descriptions and structural characteristics are not consistent. For example, both "Wu De" and "Wu Xiangfu" are married to "Qian Wenxin", thus they point to the same person with high probability.

1) Similarity Calculation Between Attributes. We divided the attributes in the database into three categories. The first type of attribute is named positive attributes, which are considered to be a relatively important class of attributes and are given the highest weight scores. For example, "place of birth" and "date of birth". The second type of attribute is called semi-positive attributes, which are given the second-highest weight score, such as "occupation" and " educational level". The third type of attribute is negative attributes. The negative attributes can exercise a veto right. That is, if two persons have a low score on the negative attributes, then they can be regarded as not similar directly. For example, two entities would not refer to the same person if they are of different genders and ethnicities. Then similarities between attributes are measured by the edit distance. Specifically, the edit distance similarity is usually calculated via the following formula [17].

$$sim\ (a_1, a_2) = \ 1 - \frac{d(a_1, a_2)}{\max(l(a_1), l(a_2))}. \tag{1}$$

Where a_1 and a_2 denote the values of attributes, $sim(a_1, a_2)$ is the edit distance similarity between a_1 and a_2, $d(a_1, a_2)$ is the edit distance between a_1 and a_2, and $l(a_1), l(a_2)$ is the string length of a_1 and a_2, respectively. Finally, the similarity of attributes is shown in:

$$S_a = \begin{cases} 0, \ if\ S_{negative} < \theta \\ w_1 * S_{negative} + w_2 * S_{positive} + w_3 * S_{semi-positive}, \ if S_{negative} \geq \theta \end{cases}. \tag{2}$$

Where S_a, $S_{negative}$, $S_{positive}$, $S_{semi-positive}$ represent the similarity of attributes, the similarity of negative attributes, the similarity of positive attributes, and the similarity of semi-positive attributes, respectively. And θ represents a threshold, w_1, w_2, and w_2 are the weight parameters.

2) Similarity Calculation Between Relations. Similarity calculation between relations refers to compare similarity between two persons based on their kinship. Specifically, as stated in Eq. (3) [30], the relational similarity between two persons S_r is equal to the ratio of the number of kinships shared by two persons to the sum of the kinship of the two persons.

$$S_r = \frac{|R(p_1) \cap R(p_2)|}{|R(p_1) \cup R(p_2)|}\ . \tag{3}$$

Where $R(p_1)$ and $R(p_2)$ denote the number of relations of p_1 and the number of relations of p_2, respectively. Then we can obtain the total similarity between two persons.

$$S_t = \ w * S_a + (1 - w) * S_r. \tag{4}$$

Where w represents the weight factor. The process of genealogical personal fusion is detailed in Algorithm 1.

4.4 Genealogical Knowledge Representation

4.4.1 Challenges of Genealogical Knowledge Representation

Challenge 1. The person may show different information in different genealogies because of privacy protection and

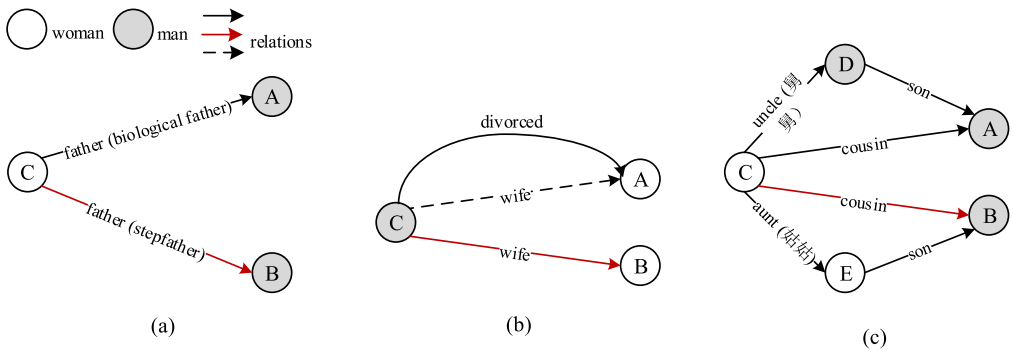


Fig. 5. Complex semantics of kinship between persons.

different identities. Moreover, the person may have different identities in different genealogies, and the descriptions of these identities are different. For example, the person Wu De is described in genealogy A and genealogy B as "Wu De Wu Taier'second son adopted to Wu Taisan" and "Qian Wenxin husband: Wu De", respectively. In other words, Wu De is an independent person in genealogy A. While in genealogy B, he is the spouse of a female person and is described in the introduction of the female person. Different descriptions of Wu De are shown for different genealogical editors. Therefore, it is a challenge to represent nodes of genealogical persons for different users properly without any information loss.

Challenge 2. The complex semantics of kinship further increase the difficulty of genealogical knowledge representation. Three types of complex semantics of kinship are described in Fig. 5. (a) Similar kinships. For example, C has multiple fathers, but in fact, A is C's biological father, and B is C's stepfather. (b) The kinships change over time. For example, C first married A, then divorced, and married B. Although both A and B are C's wives, the practical meanings of them are different. (c) The meaning of the same kinship noun may be different. For example, the cousin of C may indicate the son of an uncle or an aunt.

Algorithm 1. The Process of Genealogical Personal Fusion

Require: personal list P_A in genealogy A; personal list P_B in genealogy B.

1. Calculate candidate same person list P_{AB} between P_A and P_B according to Table 4.
2. Initialize the list P.
3. **for** i, j in P_{AB}:
4. Calculate attribute similarity S_a between i and j according to Eq. (2).
5. Calculate relation similarity S_r between i and j according to Eq. (3).
6. Calculate total similarity S_t between i and j according to Eq. (4).
7. **If** $S_t >$ Threshold β:
8. Append (i, j) to P.
9. **end if**
10. **end For**
11. Return P.

4.4.2 Our Proposed Method

We introduce the three-layer structure of "knowledge element"-"knowledge unit"-"knowledge graph" to represent genealogical knowledge. Compared with the traditional knowledge graph representation, this method is improved in three aspects: 1) The granularity of nodes in the graph structure is various including person, user, genealogy, and hypernode, which can support diversified knowledge retrieval. 2) The knowledge description ability is improved because complex semantics can be expressed through the association between knowledge element, knowledge unit, and knowledge graph. For example, the association between hypernodes and the nodes of persons can tackle the problem of how to present different information of one person in different genealogies. The hypernodes are associated with the nodes of persons through "synchronize" and "update" operations. A user can query the personal node which does not mean that he can also query the hypernode. In other words, we can present different information of a person in different genealogies for different users based on the unit of hypernode. In addition, the complex kinships between persons can be demonstrated effectively by adding constraints for knowledge elements. For example, adding constraints, i.e., "start time" and "end time", to the relation "wife" can identify the state (e.g., divorced) of the relation. 3) The framework can be described by the ontology languages such as resource description framework (RDF) and resource description framework schema (RDFS). Because the proposed method of genealogical knowledge representation inherits the characteristics of semantic networks, namely, knowledge graph.

1) Three Tiers Genealogical Knowledge Model. The overall framework of genealogical knowledge representation is shown in Fig. 6. The model constructs a three tiers knowledge network of "knowledge element"-"knowledge unit"-"knowledge graph". The multi-granularity genealogical knowledge is included and the connections with the knowledge are given in the model. The formal description of the three-tier model is detailed below.

Genealogical Knowledge Element. Genealogical knowledge element is the integral unit that is composed of concepts or entities and the relations among them. The concept of knowledge element is usually used in the field of education, such as learning resources [7]. It is regarded as the integral knowledge unit used to teaching which can exist independently and can't be divided again. In this paper,

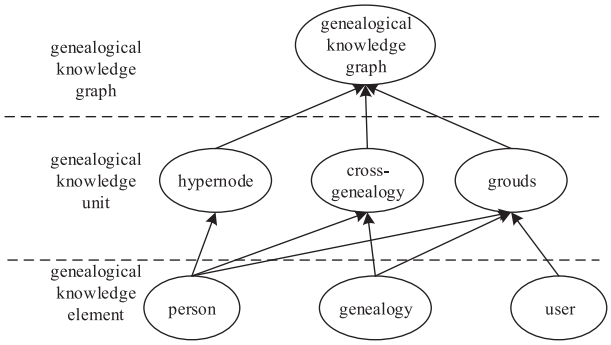


Fig. 6. Three tiers genealogical knowledge model.

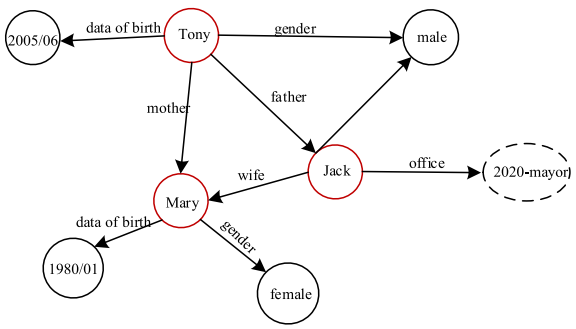


Fig. 7. The knowledge element of "person".

the knowledge element is introduced to describe the smallest complete element of genealogical knowledge. For example, an element called "person" is described in Fig. 7. Out of which a circle node denotes a concept or entity, the link between nodes represents the relation between concepts.

Genealogical Knowledge Unit. Genealogical knowledge unit refers to the compound unit consist of several knowledge elements. The knowledge unit is created to describe the complex semantics of genealogical knowledge. For example, the knowledge unit "cross-genealogy" could be used for analyzing the disputes and connections between different surnames which cannot be described by a single knowledge element.

Genealogical Knowledge Graph. Genealogical knowledge graph describes the content of genealogical knowledge in detail. It is the carrier of genealogical contents.

2) Hypernode. A hypergraph is a variant of a normal graph, and describes the multivariate relationships between objects. A hyperedge is the essential feature of a hypergraph that distinguishes it from a normal graph. A hyperedge can connect two or more vertices while an edge connects two vertices in a normal graph.

Formally, a hypergraph can be represented by a triple $G = \langle V, E, W \rangle$ [5], where V represents a set of vertices; E stands for a set of hyperedges; W represents a weight set of hyperedges. If a hyperedge $e \in E$ exists, then the degree of e can be defined as the number of vertices that e contains, namely, $|e|$.

Similarly, to describe a special type of node that contains information from more than one node, the concept of hypernode is introduced in Huapu-KG. A hypernode can be formally expressed as $N = \{v1, v2, v3, \ldots, vk\}$, where $k \geq 2$; $v1, v2, v3, \ldots, vk \in V$. That is, a hypernode can be regarded as a set of two or more vertices. The hypernode plays an important role in the construction of a genealogical knowledge graph. Specifically, a hypernode can present different information for a person in different genealogies. As shown in Fig. 8, $N_{Wu\ De}$ is a hypernode which consists of three nodes including "Wu De (1)", "Wu De (2)", and "Wu De (3)". Some users can query "Wu De (1)" while others can query "Wu De (2) and "Wu De (3)". In other words, users can query different results from the hypernode based on the user's background information.

4.5 Genealogical Knowledge Query

4.5.1 Challenges of Genealogical Knowledge Query

The one of important goals of the knowledge query is to display the query results more quickly and accurately. Specifically, 1) knowledge graph can be stored as a table [8], [19] or a graph [1]. Table-based storage refers to store all facts in a knowledge graph by a table, which is simple and straightforward. However, this method may suffer the shortcomings of costly operation and lower query speed because of a too large table, which would create a poor user experience. Graph-based storage can accurately reflect the internal structure of knowledge graph, which is conducive to deep knowledge mining. However, the structure of graph mode storage is complicated. The response speed would get slow when a large amount of load is connected. 2) We may obtain multiple results when querying a specific person in the genealogical knowledge graph due to the problems of namesakes. However, the results that we are most concerned about are only one or two. Therefore, how to filter or sort query results and return the best answer is another necessary work in genealogical knowledge query.

4.5.2 Our Proposed Method

We deal with genealogical knowledge queries from two perspectives including a distributed framework for

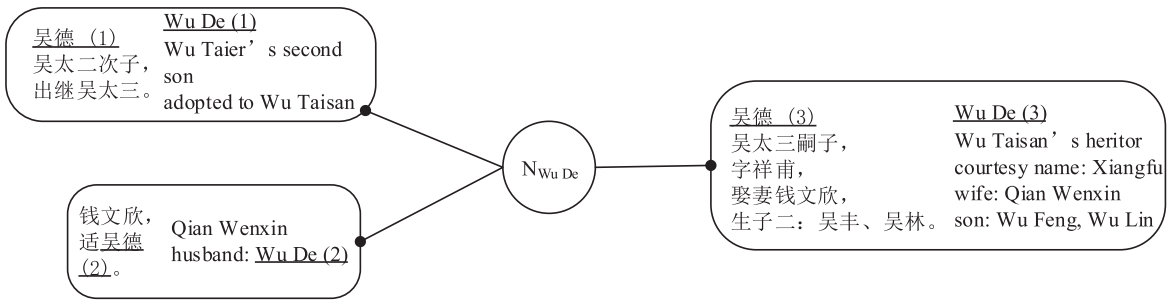


Fig. 8. An example of hypernode "$N_{Wu\ De}$".

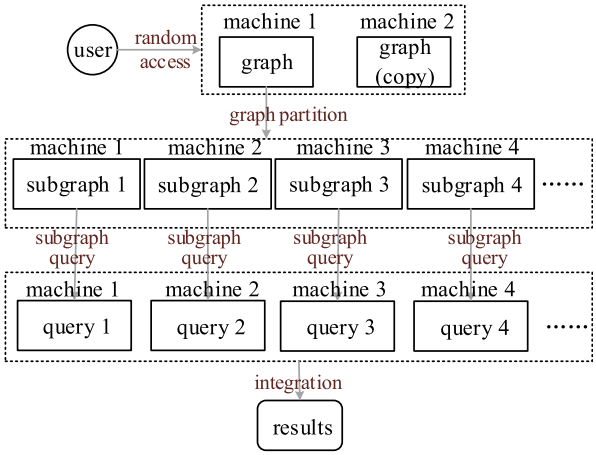


Fig. 9. A distributed framework for genealogical knowledge query.

genealogical knowledge query and personalized query of genealogical persons.

1) Distributed Framework for Genealogical Knowledge Query. A framework of "Graph database cluster + spark distributed computation" [29] established for genealogical knowledge storage and query. Specifically, the graph database system Neo4j is used to store genealogical knowledge graphs for the advantages of high performance and lightweight. Furthermore, a graph database cluster distributes the graph database on different machines and synchronizes it in real-time. We can get a faster response by distributing our requests to different machines. In addition, the Spark framework can further increase the calculating velocity and improve the user experience. The detailed flow of a query operation is shown in Fig. 9. First, the database for a genealogical knowledge graph is deployed on two machines and synchronized in real-time. The requests for accessing the database are randomly assigned to one of the machines. Second, the whole graph is divided into different subgraphs by graph partition algorithms [15] and these subgraphs are distributed to various machines for parallel operations. Third, we execute the query on the subgraphs and consolidate the subquery to get the final results.

Algorithm 2. The Process of Genealogical Personal Retrieval

Require: user: U, person: P, the genealogy which the person belongs to: G.
Genealogy PG = **QueryGenealogy** (G).
Initialization (plist).
for *p* in PG **do**:
  m = **ComparePersonSimilarity** (*p*, P).
  plist.append (m).
end for
for *p* in plist **do**:
if *p* = = U**:**
  SortList$_1$ (plist).
end if
if *p* in the genealogy managed by U:
  SortList$_2$ (plist).
end if
end for
Return plist.

TABLE 5
Statistics of Huapu-KG

knowledge		numbers	types
node	Person genealogy	15,060K+	/
	familyName	1006	/
		721	/
relation edge	Kinship socialAssociation	2942K+	485
		5449K+	471
function edge	Edit spanFamily		

2) Personalized Query of Genealogical Persons. The personal query is a typical retrieval requirement in the genealogical knowledge graph. The whole database is required to traverse when querying a person and the query speed would be seriously affected when the number of persons is too large. Therefore, we propose the personalized query of genealogical persons. On the one hand, the genealogy to which the person belongs is first located when querying a person. Then the result is obtained through compare the person that we query with other persons in the genealogy. The advantage of this method is to reduce the complexity of query operation from *O(T)* to *O(M)*, which *T* represents the total number of persons in the database and *M* denotes the number of persons in the genealogy. In general, *M* is much smaller than *T*. On the other hand, we advance the position of results that are more relevant to users for optimizing the ranking of search results. The process of the genealogical personal query is described as Algorithm 2. Steps 7-13 indicate that if the person quired equals to the user or locates in the genealogy managed by the user, it will be displayed first.

5 Evaluation and Analysis

In this section, a statistical overview of the genealogical graph is described at first. Next, evaluations about our key components of the genealogical knowledge construction are detailed.

5.1 Statistics of Our Genealogical Knowledge Graph

Genealogical knowledge graphs are stored in the form of graphs, with nodes representing entities and edges representing relations. The statistics of nodes and edges of Huapu-KG are shown in Table 5. We constructed three types of nodes including 15,060K+ persons, 1006 genealogies, and 721 family names, which include most of the Chinese family names. In addition, the edges in genealogical knowledge graphs can be basically divided into two categories: relation edges and function edges. The relation edges describe the associations between different persons including kinships and social associations.

5.2 Evaluations

5.2.1 Evaluation on Genealogical Knowledge Acquisition

We compared our method of genealogical knowledge acquisition with DSNFs [16] and fastHan [12]. The source text and the extracted results are shown in Tables 6 and 7,

TABLE 6
Source Text for Genealogical Knowledge Acquisition

吴德, 字祥甫, 吴太二次子, 出继吴太三, 娶妻钱文欣.生子二吴丰、吴林, 葬于马埠山申家巷亥山巳向有碑. Wu De, courtesy name: Xiangfu, Wu Taier's second son, adopted to Wu Taisan, married Qian Wenxin, had two sons: Wu Feng, Wu Lin, and was buried in MaBu Mountain.

respectively. From Table 7, we can observe that the relations and attributes of a person can be extracted effectively through the proposed model. However, only part of the triples can be obtained through DSNFs and only person names and places can be obtained through fastHan. This is because DSNFs and fastHan do not consider the characteristics of genealogical data.

5.2.2 Evaluation on Genealogical Knowledge Denoising

We evaluated our knowledge denoising method on the MuBuWu genealogy and the QuYangYin genealogy [22]. The experiments demonstrate that our model can discover noises in the genealogical knowledge graph effectively through defining errant relation patterns. Specifically, 46 incorrect relations were detected on the MuBuWu genealogy by utilizing the VF3 [6] algorithm and the defined schema 5 for errant relations, i.e., relation self-looping. In addition, 11 false relations on the QuYangYin genealogy were detected based on the Ullmann [9] algorithm and the defined schema 5.

5.2.3 Evaluation on Genealogical Knowledge Fusion

We evaluated our knowledge fusions method for the genealogical data in the MuBuWu genealogy and WuYueChunQiu. Specifically, the MuBuWu genealogy consists of 49285 persons and WuYuChunQiu consists of 34208 persons. Some of the persons who exist in the MuBuWu genealogy are included in WuYueChunQiu. Our goal is to align persons between the two genealogies. Note that we use sample accuracy (i.e., the proportion of correct results in the results of random sampling) and average accuracy (i.e., the average of sample accuracy) to measure the accuracy of personal alignment. Moreover, the symbol-based method for entity alignment, namely, the edit distance method [17], was selected as a comparison. As shown in Table 8, our method outperforms the baseline models because our method considers the characteristics of genealogical persons, such as the similarity of kinship between two persons.

TABLE 7
Results for Genealogical Knowledge Acquisition

method	results
DSNFs	["吴德出", "继", "吴太三"], ["吴德", "娶妻", "钱文欣"], ["吴德", "生子", "吴丰"], ["吴德", "生子", "吴林"] ["Wu De adopted", "adopted", "Wu Taisan"] (noise), ["Wu De", "married", "Qian Wenxin"], ["Wu De", "son", "Wu Feng"], ["Wu De", "son", "Wu Lin"]
fastHan	["吴德"], ["祥甫"], ["吴太二"], ["吴太三"], ["钱文欣"], ["吴丰"], ["吴林"], ["马埠山"], ["申家巷"], ["亥山"], ["巳向"] ["Wu De"], ["Xiangfu"], ["Wu Taier"], ["Wu Taisan"], ["Qian Wenxin"], ["Wu Feng"], ["Wu Lin"], ["MaBu Mountain"]
Huapu-KG	["吴德", "字", "祥甫"], ["吴德", "父亲", "吴太二"], ["吴德", "嗣父", "吴太三"], ["吴德", "妻子", "钱文欣"], ["吴德", "儿子", "吴丰"], ["吴德", "儿子", "吴林"], ["吴德", "葬于", "马埠山申家巷亥山巳向有碑"] ["Wu De", "courtesy name", "Xiangfu"], ["Wu De", "father", "Wu Taier"], ["Wu De", "adopted", "Wu Taisan"], ["Wu De", "married", "Qian Wenxin"], ["Wu De", "son", "Wu Feng"], ["Wu De", "son", "Wu Lin"], ["Wu De", "buried", "MaBu Mountain"]

TABLE 8
Results for Genealogical Knowledge Fusion

method	edit distance	Huapu-KG
sample accuracy	0.64	0.88
	0.80	0.92
	0.78	0.94
average accuracy	0.74	0.91

5.3 Challenges with the Huapu-KG Model

The proposed genealogical knowledge graph model Huapu-KG has provided a feasible solution in the life cycle of genealogical knowledge graphs based on analyzing the characteristics of genealogical data. Meanwhile, Huapu-KG still faces the following challenges.

Management of Highly Dynamic Knowledge. Huapu-KG has proposed to add constraints to dynamic knowledge, and some achievements have been made in storing a specific version of the genealogical knowledge graph at some point and in displaying the temporal characteristics of attributes. However, there is a gap in the management of highly dynamic knowledge, such as how the identities of persons have changed under dynasty changes and family's vicissitudes, as to how it has developed if a portal went out independently. To capture these changes, we not only need to understand time structure and historical knowledge, the ability of the database to store and update data should also be considered. In addition, integrity and consistency of the update process, and incremental cascading updates should be taken into consideration.

Deep inference of Knowledge. Huapu-KG automatically derives new kinships by defining basic relations and attributes between persons. It excavates hidden knowledge in genealogical knowledge graphs and promotes knowledge graph completion. But Huapu-KG lacks a deeper exploration of combining different methods to improve reasoning performance. For example, integrating additional information such as text corpus to improve the connectivity of genealogical knowledge graphs and achieve effective reasoning.

Universality of Huapu-KG. Huapu-KG provides feasible solutions for genealogical knowledge acquisition, denoising, fusion, representation, and application. Although some methods in Huapu-KG can be generalized to other

areas, it is difficult to directly applys Huapu-KG to other domains due to the particularity of genealogical data. For example, Chinese genealogy is monolingual, thus Huapu-KG may not be applicable in a multilingual knowledge system.

6 Genealogical Knowledge Graph Applications

Huapu-KG has been embedded into the China's Genealogy platform, which provides genealogy building and management services for Internet users. The applications of genealogical knowledge graphs on this platform include but are not limited to surname association, genealogical tree visualization, and social networking. We will detail these applications in the following subsections.

6.1 Surname Association

Surname association refers to establishing connections between different genealogies and mining hidden information about the origins and evolutions of surnames. This application can help expand and enrich knowledge networks, explore love and hatred histories between different genealogical persons, and search ancestors for users. The path information between persons of different surnames and defined constraints about classes and attributes in the knowledge graph provide the basis for surname association. For example, potential relationships between entities, i.e., persons, can be mined effectively through analyzing path information with the help of rule-based algorithms.

6.2 Genealogical Data Visualization

Based on Huapu-KG, complex relationships between genealogical persons are presented in multiple visualization modes. 1) Relationship diagram. The most direct kinships are shown in a person's relationship diagram. Specifically, parents, spouses, children, and siblings are included in the diagram. Additional relationships obtained by inference rules are also presented in the diagram. The relationship diagram can help us quickly find persons who are closely related to the current person. 2) Genealogical tree. Family members are linked and sorted by generation in the genealogical trees. We can learn the family origins, descendants, and basic attributes, such as "name" and "generation" /seniority of family members in the trees. 3) Personal details. Personal details are described in a "paragraph" or "table" format. In addition to the basic genealogical attributes, some specific attributes, i.e., "place of birth" and "location of burial", are also clearly described in personal details.

6.3 Social Networking

The social networking component can connect users with the same family or same interests based on knowledge graphs and form a platform for social activities. Users can share information, communicate with each other, or ask for help through this component. Relevant and popular information or topics would be preferred to the users. For example, their postings will be recommended to each other if two users belong to the same family. Specifically, the genealogical social networking focuses on the characteristics of genealogical data.

7 Conclusion

This paper proposed a genealogical knowledge graph model according to the characteristics of genealogical data. The model employs HAO intelligence to obtain professional-standard genealogical data and distinguishes properties through different granularity units. It solves the problem that a general knowledge graph model cannot be directly migrated to the genealogy domain. A real-world case study was used to verify the effectiveness of the model. Our case study has been implemented in the China's Genealogy website (https://www.zhonghuapu.com/) which currently consists of 15060000+ individuals, 721 surnames and 1006 family genealogies, and is growing. In the future, we will explore the automatic construction of larger scale genealogical knowledge graphs.

Acknowledgments

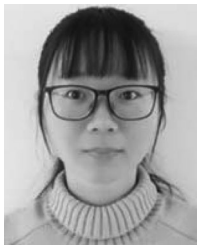
This work was supported in part by the National Key Research and Development Program of China under Grant 2016YFB1000901, in part by the Program for Changjiang Scholars and Innovative Research Team in University (PCSIRT) of the Ministry of Education under Grant IRT17R32, and in part by the Natural Science Foundation of China under Grant 91746209.

References

[1] R. Angles and C. Gutierrez, "Survey of graph database models," *ACM Comput. Surv.*, vol. 40, pp. 1–39, 2008.
[2] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, "DBpedia: A nucleus for a web of open data," in *Proc. 6th Int. Semantic Web 2nd Asian Conf. Asian Semantic Web*, 2007, pp. 722–735.
[3] A. Bordes, J. Weston, and N. Usunier, "Open question answering with weakly supervised embedding models," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discov. Databases*, 2014, pp. 165–180.
[4] A. Bordes, S. Chopra, and J. Weston, "Question answering with subgraph embeddings," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 615–620.
[5] A. Bretto, *Hypergraph Theory*. Cham, Switzerland: Springer, 2013.
[6] V. Carletti, P. Foggia, A. Saggese, and M. Vento, "Challenging the time complexity of exact subgraph isomorphism for huge and dense graphs with VF3," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 804–818, Apr. 2018.
[7] X. Chang and Q. Zheng, "Knowledge element extraction for knowledge-based learning resources organization," in *Proc. Int. Conf. Web-Based Learn.*, 2008, pp. 102–113.
[8] E. I. Chong, S. Das, G. Eadon, and J. Srinivasan, "An efficient SQL-based RDF querying scheme," in *Proc. 31st Int. Conf. Very Large Data Bases*, 2005, pp. 1216–1227.
[9] D. G. Corneil and C. C. Gotlieb, "An efficient algorithm for graph isomorphism," *J. ACM*, vol. 17, pp. 51–64, 1970.
[10] Z. Dong and Q. Dong, "HowNet–A hybrid language and knowledge resource," in *Proc. Int. Conf. Natural Lang. Process. Knowl. Eng.*, 2003, pp. 820–824.
[11] Y. Fang and J. Luo, "Modeling and implementation of kinship knowledge model based on semantics," *Comput. Tech. Develop.*, vol. 29, pp. 1–5, 2019.
[12] Z. Geng, H. Yan, X. Qiu, and X. Huang, "FastHan: A bert-based joint many-task toolkit for chinese NLP," 2020, *arXiv:2009.08633*.
[13] R. Hoffmann, C. Zhang, X. Ling, L. Zettlemoyer, and D. S. Weld, "Knowledge-based weak supervision for information extraction of overlapping relations," in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2011, pp. 541–550.

[14] G. Ji, K. Liu, S. He, and J. Zhao, "Distant supervision for relation extraction with sentence-level attention and entity descriptions," in *Proc. AAAI Conf. Artif. Intell.* 2017, pp. 3060–3066.
[15] S. Ji, C. Bu, L. Li, and X. Wu, "Local graph edge partitioning with a two-stage Heuristic method," in *Proc. IEEE 39th Int. Conf. Distrib. Comput. Syst.*, 2019, pp. 228–237.
[16] S. Jia, S. E, M. Li, and Y. Xiang, "Chinese open relation extraction and knowledge base establishment," *ACM Trans. Asian Low-Resou. Lang. Inf. Process.*, vol. 17, pp. 1–22, 2018.
[17] R. Klabunde, "Daniel Jurafsky/James H. Martin, speech and language processing," *Zeitschrift für Sprachwissenschaft*, vol. 21, pp. 134–135, 2002.
[18] C. Liang *et al.*, "Bond: Bert-assisted open-domain named entity recognition with distant supervision," in *Proc. Int. Conf. Knowl. Discov. Data Mining*, 2020, pp. 1054–1064.
[19] J. Lu *et al.*, "SOR: A practical system for ontology storage, reasoning and search," in *Proc. 33rd Int. Conf. Very Large Data Bases*, 2007, pp. 1402–1405.
[20] G. A. Miller, "WordNet: A lexical database for English," *Commun. ACM*, vol. 38, pp. 39–41, 1995.
[21] Y. Shan, C. Bu, X. Liu, S. Ji, and L. Li, "Confidence-aware negative sampling method for noisy knowledge graph embedding," in *Proc. IEEE Int. Conf. Big Knowl.*, 2018, pp. 33–40.
[22] S. Sheng, P. Zhou, and X. Wu, "CEPV: A tree structure information extraction and visualization tool for big knowledge graph," in *Proc. IEEE Int. Conf. Big Knowl.*, 2019, pp. 221–228.
[23] F. M. Suchanek, G. Kasneci, and G. Weikum, "Yago: A core of semantic knowledge," in *Proc. 16th Int. Conf. World Wide Web*, 2007, pp. 697–706.
[24] Z. Sun, W. Hu, Q. Zhang, and Y. Qu, "Bootstrapping entity alignment with knowledge graph embedding," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, 2018, pp. 4396–4402.
[25] Q. Wang, Z. Mao, B. Wang, and L. Guo, "Knowledge graph embedding: A survey of approaches and applications," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 12, pp. 2724–2743, Dec. 2017.
[26] M. Wu and X. Wu, "On big wisdom," in *Proc. IEEE Int. Conf. Data Mining*, 2019, pp. 1–8.
[27] X. Wu, B. Dong, X. Du, and W. Yang, "Data governance technology," *J. Softw.*, vol. 9, pp. 2830–2856, 2019.
[28] X. Wu, T. Jiang, Y. Zhu, and C. Bu, "Knowledge graph for china's genealogy," in *Proc. IEEE Int. Conf. Knowl. Graph*, 2020, pp. 529–535.
[29] X. Wu, S. Shen, T. Jiang, C. Bu, and M. Wu, "Huapu-CP: From knowledge graphs to a data central-platform," *Acta Automatica Sinica*, vol. 46, pp. 2045–2059, 2020.
[30] X. Wu, J. Li, P. Zhou, and C. Bu, "A fusion technique for fragmented genealogy data," *J. Softw.*, 2020. [Online]. Available: http://www.jos.org.cn/1000-9825/0000.htm
[31] X. Wu, X. Zhu, G. Wu, and W. Ding, "Data mining with big data," *IEEE Trans. Knowl. Data Eng.*, vol. 26, vol. 1, pp. 97–107, Jan. 2014.
[32] B. Xu *et al.*, "CN-DBpedia: A never-ending Chinese knowledge extraction system," in *Proc. Int. Conf. Ind. Eng. Other Appl. Appl. Intell. Syst.*, 2017, pp. 428–438.
[33] A. Yakushiji, Y. Miyao, T. Ohta, Y. Tateisi, and J. I. Tsujii, "Automatic construction of predicate-argument structure patterns for biomedical information extraction," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2006, pp. 284–292.

Xindong Wu (Fellow, IEEE) received the bachelor's and master's degrees in computer science from the Hefei University of Technology, China, and the PhD degree in artificial intelligence from the University of Edinburgh, Britain, in 1993. He is currently the director of and a professor with the Key Laboratory of Knowledge Engineering with Big Data, Hefei University of Technology, Ministry of Education, Hefei, China, and the president of the Mininglamp Academy of Sciences, Mininglamp Technology, China. His research interests include data mining and knowledge-based systems. He is a Fellow of the AAAS.



Tingting Jiang is currently working toward the PhD degree at the Hefei University of Technology, Hefei, China. Her research interests include knowledge graph, knowledge graph embedding, and entity alignment.

Yi Zhu received the BS degree from Anhui University, the MS degree from the University of Science and Technology of China, and the PhD degree from the Hefei University of Technology. He is currently an assistant professor with the School of Information Engineering, Yangzhou University, China. His research interests include data mining, knowledge engineering, and recommendation systems.

Chenyang Bu received the BE degree from the Hefei University of Technology, Hefei, China, in 2012, and the PhD degree from the University of Science and Technology of China in 2017. He is currently an assistant professor with the Hefei University of Technology. His research interests include knowledge graph embedding for dynamic data, and evolutionary dynamic optimization and applications.