Postprint: Knowledge Graph-Based Question Answering Service Framework for Red Historical Figures

Authors: Zhang Yunzhong, Guo Dong, Wang Yage, Sun Ping

Date: 2023-04-01T00:00:00+00:00

Abstract

[Purpose/Significance] Knowledge graphs have emerged as a new paradigm for knowledge organization in public digital cultural resources. Leveraging knowledge graph technology to empower question-answering services for red historical figures and enhance user interaction experience holds significant importance for the development and utilization of red historical resources. [Method/Process] Based on a review of related research on the organization of digital resources for historical figures and knowledge-based question-answering systems, we constructed a knowledge graph Schema and KBQA architecture for red historical figures, and developed a question-answering model through five stages: data acquisition, knowledge extraction, knowledge fusion, graph generation, and knowledge question-answering. An empirical study was conducted using digital resources of historical figures from Old Shanghai University. [Results/Conclusion] The knowledge question-answering service architecture designed in this paper demonstrates superiority in semi-automatic graph construction, knowledge reasoning, and intelligent interaction for red historical figure digital resources, thereby enhancing the user knowledge service experience.

Full Text

Preamble

Knowledge graphs have emerged as a new paradigm for organizing public digital cultural resources. Leveraging knowledge graph technology to empower knowledge question-and-answer (Q&A) services for red historical figures and enhance user interaction experience holds significant importance for the development and utilization of red historical resources. This study constructs a knowledge graph schema and KBQA (Knowledge Base Question Answering) architecture

for red historical figures based on a review of existing research on digital resource organization for historical figures and knowledge Q&A systems. The framework encompasses five key stages: data acquisition, knowledge extraction, knowledge fusion, graph generation, and knowledge Q&A, with an empirical study conducted using digital resources of historical figures from the old Shanghai University.

Red culture, with its distinct Chinese style and characteristics, has generated and preserved a vast array of red cultural resources in various forms throughout the long historical processes of revolution, construction, and development. Red historical figures, as the creators and disseminators of red culture, serve as the primary carriers for presenting and exhibiting red cultural resource content. In recent years, General Secretary Xi Jinping has repeatedly emphasized the importance of "making good use of red resources, carrying forward red traditions, and passing on red genes." The digital organization, management, and development of red cultural resources, particularly those related to red historical figures, carry important theoretical value and practical significance for promoting these figures and the spirit of red culture. The General Office of the Communist Party of China Central Committee and the State Council issued the "Opinions on Implementing the Revolutionary Cultural Relics Protection and Utilization Project (2018–2022)" [1], which emphasizes the appropriate use of modern technology to reconstruct and excavate knowledge associations within red historical figure resources. This not only makes it possible to display various types of information content centered on the experiences and relationships of red historical figures but also greatly enriches knowledge discovery services for these resources, thereby deepening the development and utilization of digital resources related to red historical figures.

Therefore, this study employs knowledge graph technology combined with a knowledge Q&A service framework to explore new methods for organizing, managing, and developing digital resources of red historical figures, aiming to achieve the publication of linked data on red historical figures. This will improve the data infrastructure related to red historical figures and lay a foundation for the construction and development of red tourism sites such as party history museums and museums, as well as cultural and creative products.

Research Status

This study's approach to organizing, managing, and developing digital resources of red historical figures is built upon two core issues: What is the current state of research on digital resource organization and knowledge services for historical figures? What is the current state of methodological research on Q&A services based on knowledge graphs? The following review addresses these two key points.

1.1 Research Status of Digital Resource Organization and Knowledge Services for Historical Figures

Recent research in the field of digital resource organization and knowledge services for historical figures primarily includes three aspects: First, the construction of historical figure databases. Representative databases include the ancient figure relationship database "China Biographical Database (CBDB)" [2], the "Hunan Modern Figures Resource Database" under the National Cultural Information Resources Sharing Project of the Ministry of Culture [3], and the "Li Dazhao Special Database" of Hebei red historical and cultural resources [4]. These databases provide experience for research on conceptual classification, hierarchical structure, and content selection for historical figure databases. Second, data expansion and evolution based on historical figure databases, represented by academic celebrity knowledge models based on RDF formal description [5], the Linked Data Platform for Chinese historical figure biographical data (CBDB-LD) [6], and historical figure relationship networks in CBDB [7], which lay the foundation for the formal expression and semantic association of historical figure resource data. Third, knowledge services represented by visual display of figure relationships and knowledge Q&A, including visualizations of academic master-disciple relationships in the Song Dynasty [8], historical figure entity relationship visualization systems [9], and intelligent Q&A systems for Chinese historical figure knowledge [10]. These studies broaden the research 思路 (thinking) on knowledge services for historical figures. All these studies take "historical figure digital resources" as the research object, advancing research on knowledge organization and services for historical figure digital resources and providing data and technical foundations for this study.

1.2 Research Status of Knowledge Graph-Based Q&A Service Methodologies

Knowledge graph-based Q&A services represent a hot topic in the field, with implementation methods roughly divided into four categories according to different domain knowledge content: The first category is template-matching-based Q&A methods, which rely on preset SPARQL templates [11] and select templates to generate answers based on question types. Representative services include disease Q&A systems [12] and investment Q&A systems [13]. The second category is semantic parsing-based Q&A methods, which parse natural language questions to return corresponding results, such as intelligent Q&A systems for Chinese historical figure knowledge and knowledge association and intelligent Q&A systems for museum collection resources [14]. The third category is deep learning-based Q&A methods, which optimize Q&A models through neural networks and other technologies. Representative studies include Q&A models built with LSTM neural networks [15] and intelligent Q&A systems for classical poetry knowledge graphs based on BERT and BiLSTM-CRF [16]. The fourth category is knowledge reasoning-based intelligent Q&A methods, which obtain implicit knowledge in knowledge graphs through path reasoning calculations,

such as knowledge Q&A systems based on multimodal information cyclic reasoning [17] and knowledge reasoning frameworks adopting MHRP [18]. These studies elaborate on the construction of knowledge graph-based Q&A services from multiple perspectives and provide valuable references for this research.

In summary, although existing research on the organization, management, and development of historical figure digital resources has achieved rich results, studies focusing on the construction of knowledge graph-based Q&A systems still have shortcomings: First, data sources for knowledge graph construction are primarily structured data, with few attempts to construct graphs from semi-structured data. Second, schema design for historical figure knowledge graphs inadequately matches user needs. Third, knowledge graph-based Q&A service architectures need optimization, particularly regarding intent recognition and knowledge reasoning methods. This study addresses these deficiencies by combining the two key aspects of red historical figure knowledge Q&A services.

Framework Design for Red Historical Figure Knowledge Q&A Service Based on Knowledge Graph

2.1 Two Key Issues in Red Historical Figure Knowledge Q&A Service

2.1.1 Knowledge Base Design: Schema for Red Historical Figure Knowledge Graph This study primarily employs an experience design method to design the red historical figure knowledge base, aiming to truly match the core needs of user knowledge Q&A services with the design of the red historical figure knowledge graph to enhance user interaction experience. The research recruited twenty enthusiasts of red historical figures and selected over 50 representative documents on red historical figures from data sources such as Baidu Baike, databases, and WeChat official accounts related to red literature. Through processes including material reading, interest question extraction, and question classification and focus, their Q&A needs regarding historical figures were concentrated into five aspects: basic information, revolutionary 履历 (experience), works and writings, social relationships, and archival resources. Based on these needs and combined with the information content of existing historical figure databases, the study extracted 12 main entity types, 23 relationship types, and 4 key attributes using a top-down approach, and designed the schema for the red historical figure knowledge graph, shown in Figure 1 [Figure 1: see original paper].

The schema elements of the red historical figure knowledge graph mainly cover three categories: entities, attributes, and relationships. Entities primarily reveal objective individuals related to red historical figures, such as person names, representative work titles, event names, and location names. Attributes provide structured descriptions of entity connotations, such as event backgrounds and impacts. Relationships mainly reveal certain connections between entities, such as “master-disciple/spouse/classmate/relative” relationships between “persons” and “persons.” Through the entity-attribute-description and entity-relationship-

entity triple frameworks, the network associations between knowledge nodes of red historical figures can be established and formally expressed using RDF data format. It should be emphasized that the schema design principle for the red historical figure knowledge graph is simplicity and effectiveness; information that cannot be directly obtained through triple data can be achieved through knowledge reasoning.

Taking the dashed-line relationships shown in Figure 1 as an example, data mining of revolutionary experiences in the knowledge graph can analyze implicit connections between figures and events or between figures and place names. Multi-step calculation of relationships between different figures in the graph can infer certain implicit relationships between figures. These are new "knowledge" discovered from known facts through mining, analysis, and reasoning of existing knowledge graph data.

Additionally, the schema design of the red historical figure knowledge graph should follow linked open data principles and minimize data constraints to facilitate integration with external new data, providing a conceptual foundation for data augmentation and long-path reasoning. Currently, the schema mainly covers multi-dimensional information including basic information, revolutionary experience, works and writings, social relationships, and archival resources of red historical figures. The design can be dynamically expanded as needed according to changing requirements and time. Newly added entities, relationships, and attributes only need to be linked to the existing schema according to linked data standards. The schema construction of the red historical figure knowledge graph provides underlying data infrastructure support for the knowledge organization of red historical figures, facilitating the construction of the knowledge graph and providing application support for the Q&A agent in the KBQA framework design.

2.1.2 Q&A Service Architecture: KBQA Architecture Design and Operating Mechanism The KBQA architecture is a typical form of knowledge Q&A service architecture with advantages of strong interpretability, simple deployment, and rapid implementation. The red historical figure knowledge Q&A service framework designed in this study also adopts this architecture, mainly comprising four elements: question, Q&A agent, knowledge base, and answer, as shown in Figure 2 [Figure 2: see original paper]. The question module's task is to extract natural language statements containing specific questioning intentions input by users in the chat box page through the flask framework and pass them to the Q&A agent. The Q&A agent, as the core processing framework of the Q&A service, encompasses a series of processing procedures from question recognition to answer generation, with generated answers fed back to the answer module through the flask framework to respond to questions with specific intentions. Answers are returned to users through the web chat box after processing by the Q&A core components. The knowledge base is a knowledge graph stored using NEO4J, providing data support for the overall architecture.

In terms of operating mechanism, the Q&A interaction mainly takes the form of one question and one answer in a chat box, but its core processes occur within the "Q&A agent," which mainly includes three components: natural language understanding, knowledge graph querying, and natural language generation.

(1) Natural Language Understanding. The goal of natural language understanding is to convert text into semantic representations that can be processed by machines. In this study, it mainly refers to the Q&A system's need to recognize the Q&A intent contained in user statements and convert it into corresponding query statements. First, the natural language text input by users needs to be preprocessed, mainly including automatic word segmentation, part-of-speech tagging, and stop word removal. Since computers cannot directly process text-format data, text needs to be converted into vectors. Entity recognition refers to automatically extracting words with specific meanings from user statements, such as person names and place names. Intent recognition refers to quickly judging users' true intentions based on direct or indirect information provided by users and mapping dialogue intentions to specific question types. Intent recognition itself is a classification problem, with commonly used methods including fuzzy matching-based and deep learning neural network model-based approaches.

(2) Knowledge Graph Querying. After judging the core information input by users and analyzing the questions they want to ask, the intent recognition algorithm model can automatically generate Cypher-format database query expressions to perform attribute queries, relationship queries, or knowledge reasoning in the graph database. Some knowledge query examples are shown in Table 1 .

(3) Natural Language Generation. The task of natural language generation is to reorganize language based on a correct understanding of user intentions and query results from the knowledge graph, and answer users in fluent, coherent, and understandable sentences. Natural language generation methods typically include retrieval-based and generative approaches. The former retrieves corresponding answers from the knowledge base according to intent categories and uses different rule templates to complete sentence processing and generation, with the advantage of generating relatively accurate answers but the defect of being unable to return answers when question intent cannot be recognized. The latter trains neural network models such as seq2seq, attention+BILSTM, and BERT on large amounts of labeled data to directly return answer sentences end-to-end, with the advantage of always providing answers with diverse sentence structures, but answer accuracy depends on machine learning effectiveness. Considering that generative methods have poor answer rigor and uncertainty, making them unsuitable for red historical figure knowledge Q&A, this study adopts retrieval-based methods for answer generation.

It should be noted that the implementation of knowledge reasoning in this

study combines two algorithms: Path Ranking Algorithm (PathRanking) and Translating Embedding (TransE). The TransE algorithm maps entities and relationships in the knowledge graph to low-dimensional dense space, transforming PathRanking reasoning into operations between vectors or matrices associated with entities and relationships. This operational cost is much smaller than traditional relationship path reasoning, thus significantly improving reasoning efficiency. At the same time, by mining semantic information of entities, new graph relationship paths can be constructed to help discover implicit knowledge.

2.2 Model and Process

Based on the characteristics of red historical figure digital resources and general knowledge graph construction methods, this study establishes a knowledge Q&A model for red historical figure digital resources following the principles of simplicity, scientificity, and effectiveness, as shown in Figure 3 [Figure 3: see original paper].

2.2.1 Data Acquisition Red historical figure digital resources come from relatively broad sources, typically including structured data, semi-structured data, and unstructured data. Among existing red historical figure digital resources, complete and usable structured data is relatively scarce, with most being semi-structured and unstructured data. Generally, unstructured and semi-structured data can be transformed into formatted structures through data extraction technologies and tools. This study primarily uses Python web crawlers and regular expressions to automatically extract semi-structured data from WeChat official accounts and Baidu Baike, converting the extracted data into structured JSON format to provide a foundation for the next step of knowledge extraction.

2.2.2 Knowledge Extraction Knowledge extraction refers to processing data from different sources and structures, extracting required information to form knowledge, and storing it in a certain format. According to the information needs of the red historical figure knowledge graph schema model constructed earlier, this study performs knowledge extraction on red historical figure digital resources, mainly including entity extraction, attribute extraction, and relationship extraction. Entity extraction identifies entities with specific meanings, mainly including person names, major events, representative works, and place names. Attribute extraction typically extracts attribute descriptions of entities such as persons, works, or events. Relationship extraction is usually conducted in the form of triples, responsible for extracting relationships between entities and forming knowledge networks. This study mainly uses a model combining neural networks and rules for relationship extraction. The advantage of model fusion is that it can maximally parse semi-structured and unstructured data from different sources, laying a data foundation for knowledge graph construction.

2.2.3 Knowledge Fusion After knowledge extraction, knowledge fusion methods need to be adopted to integrate the extraction results. By merging pseudonyms, aliases, and place name appellations that exist in red historical figure digital resources, entity disambiguation is completed to provide underlying support for subsequent knowledge graph reasoning. Text similarity calculation is one of the common methods for knowledge fusion, which merges entities with text similarity above a certain threshold. This study compared several common short text similarity algorithms including Jaccard similarity, edit distance, Euclidean distance, simhash algorithm, cosine similarity, and TF-IDF. Based on the principle of simplicity and effectiveness, and combined with the characteristics of entities related to red historical figure resources, this study ultimately selected entity string similarity calculation schemes—weighted edit distance algorithm and TF-IDF algorithm.

2.2.4 Graph Generation Entities, attributes, and relationships after knowledge fusion can be represented using RDF triples. The RDF triple resource description framework can effectively reveal connections between data, with serialization methods mainly including RDF/XML, N-Triples, Turtle, RDFa, and JSON-LD. The solution adopted in this paper is to store triple data visually in key-value pairs through JSON-LD, and then use the py2neo third-party library in Python to store triple knowledge into the NEO4J graph database. The advantage of this solution is fast response, strong compatibility, and easy implementation.

2.2.5 Knowledge Q&A After the knowledge graph is constructed, intelligent Q&A for red historical figure knowledge can be implemented on this basis. This study mainly achieves Q&A through three main steps: natural language understanding, knowledge querying, and natural language generation. In the Q&A design process, the study draws on the ideas of intent recognition and slot filling in multi-turn Q&A, combining machine learning algorithms to complete natural language recognition functions. After automatically converting intent recognition results into Cypher expressions, knowledge queries are completed in NEO4J by calling Python's py2neo library, and finally results are parsed and answers are generated.

Empirical Study: Knowledge Q&A for Old Shanghai University Red Historical Figures

3.1 Object Selection

This study selects digital resources of historical figures from the old Shanghai University as a case for empirical research. Historically, the active period of old Shanghai University historical figures was mostly in the 1920s, belonging to early red historical figures with outstanding historical contributions closely related to Shanghai, the birthplace of red historical activities. From a data perspective,

the research team has been collecting and organizing materials on old Shanghai University historical figures since 2014 and established the official account "Shanghai University Stories," which promotes knowledge of 52 old Shanghai University historical figures through standardized data sections in unstructured data format. Knowledge from Baidu Baike, as semi-structured data, can serve as supplementary data sources for this study. In summary, considering both historical and data perspectives, selecting old Shanghai University historical figures as a case for empirical research on the red historical figure knowledge Q&A service framework has certain representativeness and operability.

3.2 Key Steps

3.2.1 Data Acquisition for Old Shanghai University Red Historical Figures Through web crawling of 52 figure-themed posts published on the "Shanghai University Stories" official account, 146 related articles were obtained, with partial crawler programs shown in Figure 4 [Figure 4: see original paper]. To make the content more complete and comprehensive, this study also obtained some basic information and revolutionary experience descriptions from Baidu Baike profiles as supplements.

3.2.2 Knowledge Extraction for Old Shanghai University Red Historical Figures Given that the selected case lacks a proprietary named entity annotation dataset for old Shanghai University red historical figures, this study adopts a combination of rules and deep learning neural network models for knowledge extraction. Entity extraction is mainly achieved through crawler analysis and automatic extraction of proper nouns using Jieba word segmentation. Attribute and relationship extraction is mainly conducted through triple relationships, including automatic extraction based on H5 tags of semi-structured data during web crawling and automatic extraction of paragraph and chapter unstructured data using the Jiagu deep learning neural network open-source model. Based on models such as BILSTM and trained on large-scale Chinese corpora, Jiagu can perform knowledge extraction by calling relevant functions in Python third-party libraries without additional annotation due to the existence of pre-trained models. For example, when inputting "Qu Qiubai was born on January 29, 1899, in Changzhou, Jiangsu, originally named Shuang, later changed to Qu Shuang and Qu Shuang, styled Qiubai, one of the early main leaders of the Communist Party of China," the model automatically extracts and lists the triple relationships contained in the sentence, with results shown in Figure 5 [Figure 5: see original paper].

Using the above methods, this study semi-automatically extracted 531 entities, 67 attributes, and 421 triples, with subsequent manual proofreading and supplementation adding 52 entities, 2 attributes, and 39 triples. Specific knowledge extraction results are shown in Table 2 .

3.2.3 Entity Fusion for Old Shanghai University Red Historical Figures This study uses weighted edit distance algorithm and TF-IDF algorithm to calculate entity string similarity. Edit distance (Levenshtein Distance) is a string metric in NLP that calculates the difference between two strings, characterized by using dynamic programming to compare text structures, offering high speed and accuracy for short text similarity calculation. The Levenshtein Distance between two red historical figure digital resource entity strings a and b can be represented as lev_a,b(|a|,|b|), where |a| and |b| correspond to the lengths of a and b. Through matrix calculation and iterative loops, the edit distance between two entity strings can be obtained and recorded as L_{Distance}.

TF-IDF is a statistical method where word vectors extracted through TF-IDF algorithm can well reflect differences between entity strings. After extracting word vectors of two entity strings, their similarity value sim(a,b) can be calculated through vector cosine formula, recorded as T_{Distance}. Through experimental tuning, the final similarity between two entity strings is calculated as shown in formula (3): similarity = $0.6L_\{Distance\} + 0.4$T_{Distance}. When the similarity value is greater than 0.85, strings a and b are considered the same entity and can be merged using Cypher's merge function; otherwise, they are considered different entities. Through these knowledge fusion steps, this study merged 236 entities, completing the integration of multi-source entity information, which facilitates the system's inclusion of data from various channels to enrich triple knowledge and effectively guarantees the accuracy of subsequent reasoning and Q&A.

3.2.4 Graph Generation for Old Shanghai University Red Historical Figures After organizing the triple data, using Python's third-party library py2neo and Cypher statements, triple data can be automatically imported into the NEO4J graph database. Relying on NEO4J's visualization function, a visual knowledge graph of old Shanghai University red historical figures can be displayed. The knowledge graph contains 347 entities, 69 attributes, and 460 relationships, with a partial visualization interface shown in Figure 6 [Figure 6: see original paper].

From Figure 6, triple information of some old Shanghai University historical figures can be intuitively seen. For example, Shi Cun tong, Chen Wangdao, Feng Zikai, and Zhang Chongwen are all from Zhejiang Province; Li Shutong is the teacher of Feng Zikai and Wu Mengfei; One of Chen Wangdao's representative works is the translation of the Communist Manifesto, created in 1920, and content excerpts can be seen from the entity attributes of the representative work; Yuanyuan Hall is a relic related to Feng Zikai, located in Zhejiang.

3.2.5 Knowledge Q&A for Old Shanghai University Red Historical Figures

(1) Text Preprocessing. This case uses the open-source tool Jieba for automatic word segmentation and part-of-speech tagging. Through Jieba's

custom dictionary function, entities in the old Shanghai University historical figure knowledge graph are created into Jieba's custom dictionary using Python, enabling automatic extraction of entities contained in user questions and laying a foundation for intent recognition.

(2) Intent Recognition. This case constructs the main model for intent recognition using slot filling and Naive Bayes algorithm, supplemented by fuzzy matching algorithms for secondary recognition of statements with unclear intentions. "Slot filling + machine learning classification algorithm" is a method that can accurately identify user question categories. Slots correspond to information needed to be filled to transform preliminary user intentions into clear user instructions during dialogue [19]. A slot corresponds to a piece of information required in one Q&A processing, and answering complete questions usually requires constructing chain slots based on entities, attributes, and relationships and filling them accordingly. Taking the question "What unit did Qu Qiubai work in in 1937?" as an example, chain slots can be constructed as follows: after entity recognition, it automatically fills to {'entity_{num}': [1], 'entity': ['Qu Qiubai'], 'quality_{deep}': ['True'], 'quality': ['position'], 'intent': []}.

After slot filling is completed, the filled slot attributes need to be mapped to specific questions, a function assisted by the Naive Bayes algorithm. For results of different slot fillings, the Naive Bayes algorithm learns the joint probability distribution of inputs and outputs of the intent recognition model and calculates the output with the maximum posterior probability, which is the most likely dialogue intention. If user statements contain insufficient information to fill enough slots for intent classification, fuzzy matching is used as an auxiliary model. This case uses the sum of edit distance and cosine similarity algorithms weighted as the core step for short text similarity calculation, taking the highest probability value in descending order as the most likely Q&A intention.

3.3 Results Demonstration

3.3.1 System Structure The prototype framework of the intelligent Q&A system built in this case is shown in Figure 7 [Figure 7: see original paper], mainly including data acquisition, data processing, graph generation, intent recognition, SQL statement generation, answer querying, and front-end/back-end Flask interaction. The system prototype's front-end uses H5 to build a web chat box-style service; the back-end uses Python to implement the underlying functions of intelligent recognition, querying, and generation; front-end/back-end interaction adopts the popular Flask web framework, which features a relatively simple core structure but strong extensibility and compatibility, thus enabling rapid implementation of a web intelligent chat service.

3.3.2 Q&A Examples The Q&A examples in this case mainly involve typical questions about basic information, social relationships, revolutionary experiences, works and writings, digital archives, and intelligent reasoning of red

historical figures. Figure 8 [Figure 8: see original paper] shows the interaction process of the Q&A system. Basic information Q&A demonstrates questions about the birthplace, birth year, and personal achievements of Qu Qiubai, Wu Mengfei, and Ma Ning. Social relationship Q&A shows directed social relationship questions such as "Who is Xu Xinying's teacher?" and "Who introduced Wang Huanxin to the Party?", as well as traversal social relationship questions like "What are Wu Mengfei's social relationships?". Revolutionary experience Q&A demonstrates questions about the revolutionary 履历 (experience) of three old Shanghai University historical figures—Ke Bonian, Shi Cuntong, and Feng Zikai—in different years. Works and writings Q&A shows questions about Chen Wangdao and his works, such as "What are Chen Wangdao's representative works?", "What type of work is the Communist Manifesto?", and "Recommend me some wonderful excerpts from the Communist Manifesto". Digital archives Q&A demonstrates questions about some figures' digital archive names, multimedia links, storage addresses, and locations. Intelligent reasoning Q&A shows answers to questions like "Which old Shanghai University historical figures are related to the May Thirtieth Movement?", "Which old Shanghai University figures participated in the founding ceremony?", and "Who is from Zhejiang?".

3.3.3 Q&A Testing This study conducted system function testing on the implemented Q&A service in terms of answer accuracy and response speed. The research collected over 200 popular questions from old Shanghai University historical figure knowledge Q&A enthusiasts, and through manual proofreading, classification, and supplementation, created a test dataset of 150 Q&A pairs (including questions and standard answers) covering six themes: basic information, social relationships, revolutionary experience, works and writings, digital archives, and intelligent reasoning. Testing showed that the system's average accuracy rate reached 90.6%. Experiments demonstrate that the automatic Q&A system developed in this study can answer most questions accurately. Specific test results are shown in Table 3 .

Concurrently, using Apache JMeter tool to test the Q&A system's speed performance with 200 concurrent users, results are shown in Figure 9 [Figure 9: see original paper]. The average dialogue service delay is 0.16 seconds, which can meet scenarios where over 100 people simultaneously send requests to the Q&A system and receive quick responses. However, for applications with broader regions and larger user numbers, the delay reaches over 0.5 seconds, making it unable to respond quickly, indicating that performance still needs improvement.

3.4 Result Analysis and Discussion

This study uses knowledge graphs to describe, organize, and associate digital resources related to old Shanghai University historical figures and implements an intelligent Q&A system prototype based on knowledge graphs on this foundation. The superiority of this system is reflected in the following points: (1) It achieves visual display and innovates resource organization methods, pro-

viding technical solutions for integrating red historical figure resources from semi-structured and unstructured data and constructing knowledge graphs, and realizing storage and visualization of key information resources of red historical figures relying on graph databases. (2) It improves answer accuracy through graph-based associations. Dialogue systems based on knowledge graphs excel at solving Q&A in vertical domains, and the old Shanghai University historical figure knowledge Q&A system belongs to this category, achieving intelligent Q&A for linked data with higher accuracy than chit-chat Q&A systems. (3) It achieves intelligent interaction with good promotion and application prospects. The old Shanghai University historical figure knowledge Q&A system can be packaged as a WeChat mini-program for a broader audience, greatly improving the popularization of red historical figure digital resources, stimulating readers' learning interest, and providing a new approach for red knowledge promotion and dissemination.

The empirical case also reveals many shortcomings in the research: limited by the scope of old Shanghai University historical figures, the sample size is small; there is a lack of sufficient annotated red historical figure training sets, requiring manual proofreading to supplement automatic extraction; speed performance is insufficient when concurrency exceeds 200; and the application form of Q&A display is not diversified enough. These are issues that need to be addressed in future research.

The main innovations of this study focus on the KBQA architecture design and operating mechanism for red historical figure digital resources: Combining the characteristics of red historical figure knowledge and drawing on slot filling schemes for intent recognition in multi-turn Q&A, it proposes an intent recognition method of "slot filling + machine learning classification algorithm," improving intent recognition accuracy; Using a combination of TransE and PathRanking algorithms to achieve intelligent reasoning based on knowledge graphs.

Under the digital humanities background, the continuous maturation and in-depth application of big data and artificial intelligence technologies have changed traditional knowledge organization and service methods. Effectively utilizing these technologies will cultivate new momentum for the transformation of red digital resource knowledge organization and services. The knowledge Q&A system for old Shanghai University historical figures implemented in this paper provides a general solution for extracting key data on red historical figures from multi-source data and constructing knowledge graphs. Based on knowledge graph construction, it explores typical applications of red historical figure knowledge Q&A systems, providing new ideas for knowledge graph technology empowering knowledge organization and service models, and offering new paths for the deep development and utilization of red digital resources.

Next steps for this research will include: on one hand, expanding the red historical figure sample set and exploring models with higher accuracy and faster speed to construct red historical figure knowledge graphs at scale, laying a solid data

infrastructure; on the other hand, adopting diversified forms such as WeChat mini-programs and APPs to broaden application channels for red digital resource knowledge services.

References

[1] General Office of the Communist Party of China Central Committee, State Council. Opinions on Implementing the Revolutionary Cultural Relics Protection and Utilization Project (2018-2022) [N]. People's Daily, 2018-07-30(001).

[2] Harvard University, Academia Sinica, Peking University. China biographical database [EB/OL]. [2021-03-03]. https://projects.iq.harvard.edu/cbdb.

[3] Zhang Wenyong. Research and Implementation of Hunan Modern Figures Database Construction [D]. Changsha: Central South University, 2013.

[4] Jin Zhijun, He Shoufeng, Hao Chunliu, et al. Research on the Excavation of Hebei Red Historical and Cultural Resources: Taking the Construction of Li Dazhao Special Database as an Example [J]. Culture Monthly, 2016(7): 114-115.

[5] Liu Ningjing, Liu Yin, Wang Moyan, et al. Academic Celebrity Knowledge Model Construction from a Digital Humanities Perspective [J]. Library and Information Service, 2019, 63(23): 113-121.

[6] Chen Tao, Liu Wei, Shan Rongrong, et al. Research on the Application of Knowledge Graphs in Digital Humanities [J]. Journal of Library Science in China, 2019, 45(6): 34-49.

[7] Pan Jun. Research on Distributed Semantic Representation of Figures for Digital Humanities: Based on CBDB Database and Ancient Books and Documents [J]. Library Journal, 2020, 39(8): 94-102.

[8] Yang Haici, Wang Jun. Construction and Visualization of Song Dynasty Academic Master-Disciple Knowledge Graph [J]. Data Analysis and Knowledge Discovery, 2019, 3(6): 109-116.

[9] Zhou Yi, Zhou Mingquan, Wang Xuesong, et al. Construction and Implementation of Historical Figure Knowledge Graph in Big Data Environment [J]. Journal of System Simulation, 2016, 28(10): 2560-2566.

[10] Shan Liang, Liu Xin. Construction of Intelligent Q&A System Based on Chinese Historical Figure Knowledge [J]. Information Research, 2019(6): 101-105.

[11] COCCO R, ATZORI M, ZANIOLO C, et al. Machine learning of SPARQL templates for question answering over Linked Spending [C]//CEUR workshop proceedings, 2019, 2400: 156-161.

[12] Li He, Liu Jiayu, Li Shiyu, et al. Research on Optimization of Automatic Q&A System Based on Disease Knowledge Graph [J]. Data Analysis and Knowledge Discovery, 2021, 5(5): 115-126.

[13] Chen Jinghao, Zeng Zhen, Li Gang. Construction of "Belt and Road" Investment Q&A System Based on Knowledge Graph [J]. Library and Information Service, 2020, 64(12): 95-105.

[14] Gao Jinsong, Fang Xiaoyin, Liu Siyang, et al. Research on Knowledge Association and Intelligent Q&A of Museum Collection Resources Based on Linked Data [J]. Information Science, 2021, 39(5): 12-20.

[15] WU W Q, ZHU Z F, LU Q, et al. Introducing external knowledge to answer questions with implicit temporal constraints over knowledge base [J]. Future Internet, 2020, 12(3): 45.

[16] Xie Xiang. Research on Intelligent Q&A Based on Classical Poetry Knowledge Graph [D]. Wuhan: Central China Normal University, 2020.

[17] YU J, ZHU Z H, WANG Y J, et al. Cross-modal knowledge reasoning for knowledge-based visual question answering [EB/OL]. [2021-08-02]. https://www.sciencedirect.com/science/article/pii/S0031320320303666?via%3Dihub.

[18] HUANG T S, LI X W, ZHAI S P, et al. Knowledge graph reasoning based on tensor decomposition and MHRP-Learning [EB/OL]. [2021-08-02]. https://downloads.hindawi.com/journals/am/2021/8880553.pdf.

[19] TANG H, DONG H J, ZHOU Q J. End-to-end masked graph-based CRF for joint slot filling and intent detection [J]. Neurocomputing, 2020, 413(6): 348-35.

Author Contributions

Zhang Yunzhong: Determined the research topic, proposed research ideas, designed the research plan, and wrote and revised the paper; Guo Dong: Designed programs, conducted experiments, and wrote and revised the initial draft; Wang Yage: Processed data and revised the paper; Sun Ping: Processed data and revised the paper.

Framework of Knowledge Q&A Service for Red Historical Figures Based on Knowledge Graph

Zhang Yunzhong, Guo Dong, Wang Yage, Sun Ping

Department of Library, Information and Archives, Shanghai University, Shanghai 200444

Abstract: [Purpose/Significance] Knowledge graph has become a new form of public digital cultural resource organization. Using knowledge graph technology to enable knowledge Q&A services for red historical figures and improve user interaction experience is of great significance to the development and utilization of red historical resources. [Method/Process] Based on reviewing related research

on digital resource organization and knowledge Q&A systems for historical figures, this paper constructed a knowledge graph schema and KBQA architecture for red historical figures, and built a Q&A model for red historical figures from five aspects: data acquisition, knowledge extraction, knowledge fusion, graph generation, and knowledge Q&A, with an empirical study conducted using digital resources of historical figures from Shanghai University (1922-1927). [Result/Conclusion] The knowledge Q&A service architecture designed in this paper has advantages in semi-automatic graph construction, knowledge reasoning, and intelligent interaction of digital resources of red historical figures, and improves user knowledge service experience.

Keywords: red historical figures; knowledge graph; question answering system; knowledge service

Note: Figure translations are in progress. See original paper for figures.