2021 10th International Conference on Educational and Information Technology (ICEIT) | 978-1-6654-2295-6/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICEIT51700.2021.9375538

"The Canton Canon" Digital Library Based on Knowledge Graph -- Taking the Revolutionary Archives of Canton in the Republic of China as an Example

Junchao Wu, Ying Jiang, Xin Chen, Lingyu Guo, Xiaotong Wei, Xiaoyan Yang
School of Management, Beijing Normal University, Zhuhai
Zhuhai, China
wujunchaoIU@outlook.com

***Abstract*—In order to more efficiently meet people's in-depth needs for knowledge acquisition and research, and to more effectively develop and use "the Canton Canon" historical archives, we designed and constructed the Canton Revolutionary History Knowledge Graph based on the revolutionary archives of Canton during the Republic of China. We associated the knowledge graph with the document corpus, transformed and combined the traditional digital library from document information management to knowledge management, and proposed a double-Layer retrieval mode of the document corpus based on the knowledge graph. At the same time, we developed "the Canton Canon" digital library based on the knowledge graph, and introduced the design and application of the digital library. It provides a great example of how to use huge archives more efficiently and build a more open digital library. What's more, it also opens up new ideas for the transformation of Chinese digital libraries from information management to knowledge management.**

Keywords—document retrieval, knowledge graph, digital library, the Canton Canon

I. INTRODUCTION

The Propaganda Department of the Canton Municipal Committee of the Communist Party of China and the Guangdong Provincial Department of Culture took the lead in organizing the compilation and publication of "the Canton Canon" [1] in 2005, which was based on the collections of Guangdong Provincial Sun Yat-sen Library and Sun Yat-Sen University Library, supplemented by collections of domestic and overseas public institutions and individuals. After years of hard work, "the Canton Canon" has now included 4064 kinds of documents, compiled into 520 volumes, each with about 850 pages. It is estimated that the scanned documents can be piled up to 30 stories high. "The Canton Canon" has made a significant contribution to the timely rescue and protection of the existing precious historical literature resources, and has achieved substantial results. Base on this, how to develop and utilize such a huge historical archives library effectively is the next problem to be solved urgently. To solve this problem, "the Canton Canon" Research Center of the Canton Library has developed "the Canton Canon" Database to provide literature browsing and search functions in the form of a digital library, which supports several search modes such as title, responsible person, and keyword full-text retrieval to match the content of interest to the user. In general, however, it relies heavily on its indexing and keyword retrieval to match relevant bibliographic information.

The catalogue index can help users to understand the structure of a library's digital resources quickly. However, in its huge amount of documentary corpus, when users want to search through its index to gain knowledge, it will be difficult for them to obtain more comprehensive information and inefficient to develop and utilise documentary resources if they do not have sufficient domain knowledge and do not know the relationships between knowledge points. Although keyword retrieval can quickly reveal bibliographic material that overlaps with content keywords, relying on the presence or absence of a single keyword as a search criterion for bibliographic material is not a good way to correlate knowledge. Too much reliance on keywords leads to a decrease in the development and utilisation of documentation resources. It is likely to lead to a situation where too many or too few documents are checked out, or they are not in line with users' needs. Therefore, the traditional "Canton Canon" Database, which is aimed at document management and focuses on keyword retrieval, is still unable to efficiently meet people's deep needs for knowledge acquisition and knowledge research.

Under these circumstances, this paper proposes to design and construct a knowledge graph of the revolutionary documents of Canton during the Republic of China, using the documents of Canton during the Republic of China as a breakthrough, to develop a digital library of "the Canton Canon" based on the knowledge graph. This digital library will directly turn historical people, events and other knowledge into managed objects, using the exploitation of complex relationships between knowledge as a means to provide functions such as knowledge browsing, knowledge retrieval, knowledge evolution and knowledge visualisation that are not possible in traditional digital libraries. In summary, this paper explores new ideas for the transformation of digital libraries from information management to knowledge management with the help of knowledge grapy theories and technology. It also seeks to provide a model for application in the field of revolutionary history of Canton during the Republican period, and to apply the section on the historical documents of the

Republican Revolution in the second phase of "the Canton Canon", which is of certain theoretical and practical significance.

II. Analysis of Current Research Situation

A. Overview of Digital Library Development

The emergence of digital libraries began in 1994 when the National Science Foundation (NSF), the Defense Advanced Research Projects Agency (DARPA) and the National Aeronautics and Space Administration (NASA) jointly launched a four-year research project on digital library [2]. Sun Chengjian [3] summarized it as an easy-to-use, mega-scale knowledge center with no time and space constraints. The service of digital library is an information management model for the future development of the Internet by transmitting digital information, such as text, images, and sound, through the Internet in a way that is guided by knowledge concepts, thereby sharing information resources.

With the application of electronic tagging technology to digital library, it can automatically identify, classify, organize information, and improve the efficiency of retrieval. However, in terms of information organization, knowledge and literature information are not well combined, and there are still cases where the integrated literature resources are not comprehensive enough or are too redundant, resulting in insufficient utilization of resources [4]. At the same time, the search efficiency is not high due to the large number of documents, duplicated contents, and multiple sources, which makes the knowledge in the massive documents cannot be fully explored and utilized by people. What's more, it is difficult to explore the intrinsic connection between the knowledge in the documents and can not fully meet the needs of users.

Digital library has become the main form of modern library. Broadly speaking, it includes the combination of modern information technology and library applications, with the stage of development of the social environment as the morphological classification standard, in the face of the transformation of the social environment from "information world → knowledge world → wisdom world", the library form is also changing from "electronic library → digital library → smart library" [5]. From the perspective of the evolution of digital library, "the Canton Canon" Database is still in a relatively backward electronic library stage, and it is in urgent need of a more advanced digital library in order to develop and utilize its historical archives more effectively.

Most modern digital libraries are transitioning from information management to knowledge management, and the Making of America Digital Library (MOA) in the United States and the Zhongshan Library in Guangdong Province in China represent the frontier of digital libraries in the world. Some digital libraries, such as CNKI, have gradually explored the direction of intelligent library in the field of user service, and have improved the intelligibility of their services by constructing user portraits and other methods. Unfortunately, although the Rising America Digital Library and the Zhongshan Library of Guangdong Province provide a variety of search and access services based on title, author, subject, and other published information and content, they rely primarily on keyword matching and do not further explore the connections between knowledge in the content. CNKI began to try to build a "knowledge network" through micro-level connections from scattered knowledge in 2002, and to develop knowledge element extraction and mining technology [6]. It takes knowledge graph technology as the main direction of current and future intelligent technology realization and launches the knowledge service strategy. However, its management of document knowledge still stays at the relation of "HowNet", and does not go deep into the specific content of the document; in addition, the development of the content knowledge of the document is not comprehensive enough.

B. Research Status of Knowledge Graph

The concept of Knowledge Graph was proposed by Google in 2012 [7] and was announced as the basis for building the next generation of intelligent search engines. Its main core is to use the relationship between entities and the attribute information of entities to solve the problem of intelligent question and answer related to entities, forming a new information retrieval mode. As the most effective way to express the relationship, it uses visualization technology to describe knowledge resources and their carriers, mining, analyzing, constructing, drawing and displaying knowledge and their interconnections [8]. It is essentially a semantic network, a graph-based data structure composed of points and edges: points are used to display semantic symbols such as "entities", and edges are the "relationships" between entities. Its ability to describe data is very powerful, fully demonstrating the ability to analyze problems from the perspective of "relationships".

Since 2015, knowledge graph technology has been widely used in practice. Especially the Internet industry, which has a large amount of data, is at the forefront of the world in the application of knowledge graph. After several years of cultivation, companies in different fields, such as medical care, finance, and customer service, have established market reputations. In foreign countries, IBM launched the "moon shot", which integrates a large number of medical documents, books and various EMRs (electronic medical records) to obtain massive and high-quality medical knowledge, and based on this knowledge, provides medical staff with auxiliary clinical decision-making and application of drug safety and other aspects. In China, Tonglian Data Company extracts relevant data of listed companies and integrates it into a knowledge graph to help enterprises or investment institutions conduct network-wide data correlation analysis, influence dissemination and prediction. In the meantime, a series of knowledge graph products and applications such as Baidu Zhixin and Sogou Knowledge Cube emerged in China. In addition, knowledge graph is penetrating into many other vertical domains.

The Chinese encyclopedic knowledge graph, zhishi.me [9], was constructed and published by Shenzhen Gowild Robotics Co., Ltd. and Southeast University. It has knowledge from Baidu, Interactive Encyclopedia and Chinese wiki encyclopedia, with tens of millions of entities and billions of relationships. Retrieving a knowledge point on this platform, we can not only find the information data of the knowledge point, but also find the category to which the knowledge point belongs and the associated knowledge point. However, these

retrieved knowledge triple data sets are limited and can not meet users'' implied knowledge needs well. For example, when searching for “Sun Yat-sen”, an user find that “Sun Yat-sen” has a related event named “The Revolution of 1911”. At this time, the triple data do not provide a good description of the background knowledge about “Sun Yat-sen” and “The Revolution of 1911”, and it is necessary to search for relevant literature information to supplement the description.

In summary, most traditional digital libraries, including “the Canton Canon” database, have a wealth of documentary information, but they are not well equipped to relate the knowledge points in the content. In contrast, knowledge graph technology, which has a strong ability to describe knowledge relationships, but lacks the ability to describe the relevant background knowledge of associated knowledge points. In this paper, we apply knowledge graph theory technology to digital libraries, organize the information of the Republic of China's Canton Revolution and design relevant retrieval modes, and construct “the Canton Canon” digital library, which has a huge essential advantage over "the Canton Canon" Database in every aspect. Table I is a specific comparison table between the two.

TABLE I. COMPARISON TABLE BETWEEN "THE CANTON CANON" DATABASE AND THE RESEARCH RESULTS OF THIS PAPER

	“the Canton Canon” Database (Traditional Digital Library)	***“the Canton Canon” Digital Library (Research Results of this paper)***
Information organization	A single digital Canton historical and cultural works of integrated document information.	The knowledge graph of Canton revolution history as the main knowledge information and its related digital documents and pictures.
retrieval entrance	It mainly relies on its catalog index and keyword retrieval.	In terms of knowledge service, it supports catalog classification retrieval, keyword fuzzy retrieval, map extension retrieval, relation retrieval, entity backtracking and space-time retrieval. In the way of document information, it supports knowledge keywords, N-triplet sets retrieval.
Retrieval model	Document retrieval mode of single keyword	The double-layer document corpus retrieval model combines the document layer and the knowledge layer for retrieval.
Information visualization	Single document image	In the aspect of knowledge service, knowledge graph visualization, entity information classification and other aspects of information display. In terms of literature information, keyword highlighting is supported, and images can also be displayed.

III. CANTON REVOLUTIONARY INFORMATION ORGANIZATION OF

The organization of the Republic of China Canton Revolutionary Information based on the knowledge graph is mainly divided into two parts: the content of the Republic of China Canton Revolutionary Knowledge Graph and the collection and transformation of information from the Republic of China Canton Revolutionary Knowledge Graph. Among them, the content of the Canton Revolutionary Knowledge Atlas during the Republic of China is to design and model the domain knowledge of the Canton Revolutionary History in the Republic of China, determine the scope of its domain content and build the norms of the ontology database to better collect and transform domain knowledge information. The collection and transformation of information on the revolutionary history of Canton during the Republic of China is a specification for the collection and storage of information on the revolutionary knowledge spectrum of the Republic of China. The semi-automatic information collection, storage and transformation of knowledge graphs are realized through related technologies such as web crawlers and data mining, so as to construct the knowledge graphs.

The knowledge graph of the Revolutionary Republic of Canton is open, which refers to the openness of the knowledge domain and the openness of the information organization. The core of solving the problem of co-construction and sharing of digital resources is the issue of metadata standards [10]; thus, we adopts the XML/RDF framework to encapsulate and integrate digital resources. The knowledge graph of the Revolutionary Republic of Canton is not only applicable to the field of Canton Revolutionary History in the Republic of China, as the width and depth of the knowledge graph are expandable and infinite. The design of knowledge in different fields can help the knowledge graph to continuously expand and be compatible with knowledge in different fields. In terms of information organization, taking the domain knowledge as the starting point, we can continue to extend it, collect related knowledge information and bibliographic data to expand and perfect the domain information, and continue to expand iteratively, which can help to extend the depth of the knowledge graph of the Revolutionary Republic of Canton. “The Canton Canon” is a vast historical series covering many aspects of social, economic, political, and military history. Using the results of this paper as a model for application, it can be gradually extended to any other topics included in “the Canton Canon”, even to the areas outside the canon, such as Hakka literature and Chaoshan literature, in addition to the revolutionary history of the Republic of China. Along the line of correlation of the knowledge graph, we continue to expand, self-improve, and self-enrich to form a large, multi-domain, open knowledge graph.

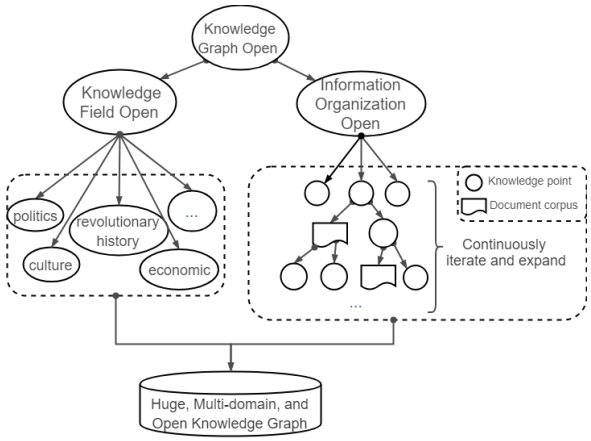

Fig. 1. An open diagram of the Knowledge graph of the Revolution in Canton in the Republic of China

A. *Main Contents of the Knowledge Graph of the Revolution in Canton in the Republic of China*

The Canton Revolutionary History Knowledge Graph of the Republic of China is composed of Knowledge Classification (Classes), Knowledge Points (Instance), Knowledge Attributes (Datatype Property), Knowledge Attribute Values (Datatype Property Value), and the Association between Knowledge Points (Object Property). These are common words in the ontology domain, and their distinction is based on the semantic meaning of the words in the domain.

The vocabulary often found in the field, which is the knowledge point, for example: “Sun Yat-sen”, “Canton Uprising”, “Huangpu Treaty”, etc. Canton's classification of knowledge in the field of revolutionary history is constructed in a top-down manner and is multi-layered, with the first level of knowledge classified into eight categories: figures, historical events, cultural relics, treaty and declaration, time, location, and historical fragments. The creation of secondary classifications based on primary classifications. For example, the characters are divided into “Communist figures”, “Kuomintang figures”, “Figures of the Chinese Peasants and Workers Democratic Party”, “Figures of China Zhi Gong Party”, “Figures of the Chinese Revolutionary Party”, Figures without party affiliation" and "Figures of other parties". When one of the other parties has a certain number of bases, a corresponding secondary class is constructed to standardize the knowledge classification, which eventually forms the storage form of a knowledge tree. Each refined class is called a “subClass”, and the refined class is called a “topClass”. The knowledge points are generally the bottom branches of the knowledge tree, which become the leaves. The specific classification knowledge tree of the knowledge base is shown in Figure 2 below.

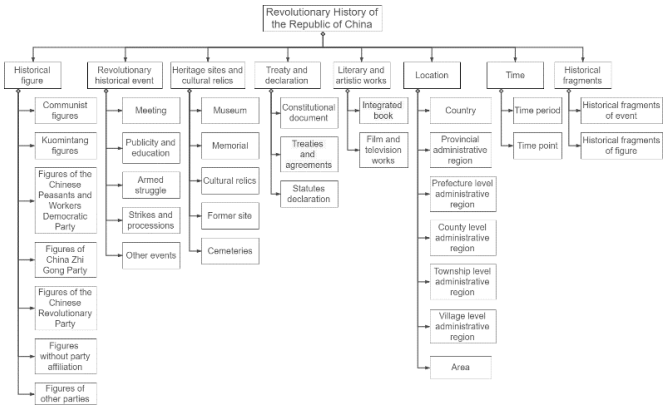

Fig. 2. Caonton Revolutionary History Field Knowledge Tree

A knowledge attribute describes an aspect of a knowledge point. As shown in Figure 3, for the knowledge point "Huanghuagang Uprising", it has knowledge attributes such as “participating parties” and “results”. For the knowledge point “Sun Yat-sen”, it has “occupation”, “graduation school”, “alias”, and other knowledge properties.

When a knowledge point is described by a knowledge attribute, the value of a knowledge attribute is assigned to that knowledge attribute to form a complete description of the knowledge point. As shown in Figure 3, for the knowledge point “Huanghuagang Uprising”, the value of the knowledge attribute of its “participating parties” is “Qing and Revolutionary Armies”, and the value of the knowledge attribute of the “ result” knowledge attribute is “Uprising suppressed”.

Knowledge attributes and knowledge attribute values characterize knowledge points, knowledge relations are also used to describe knowledge points. But knowledge relationships describe how knowledge points are related to other knowledge points. As shown in Figure 3, The relevant figure of the “Huanghuagang Uprising” is “Sun Yat-sen” is described in the knowledge graph as follows: first define a knowledge relationship “related figure”, then use a directed arrow to point from the knowledge node of “Huanghuagang Uprising” to the knowledge node of “Sun Yat-sen”, and mark this knowledge relationship as “related figure”. In this definition process, the starting point of the directed arrow is called the “subject”, and the ending point of the directed arrow is called the “object”.

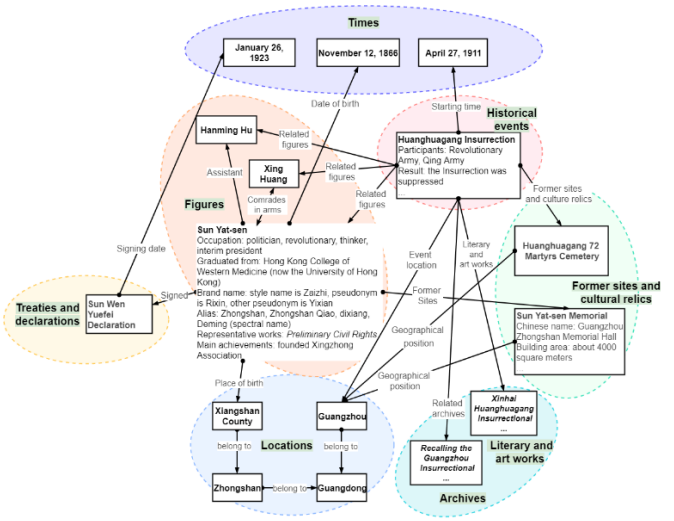

Fig. 3. Schematic diagram of knowledge map in the field of Canton revolutionary historical events

Figure 3 shows part of the knowledge graph in the field of Canton revolutionary historical events, describing the knowledge network formed by the “Huanghuagang Uprising” : Sun Yat-sen, Hu Hanmin, and Huang Xing led the Huanghuagang Uprising in Canton on April 27, 1911.The participating parties were the Qing Army and the Revolutionary Army. In the end, the uprising was suppressed. The remaining cultural relics include the Huanghuagang Seventy-two Martyrs Cemetery, and the related literature and art works and archives include the “Xinhai Huanghuagang Uprising” and “Recalling the Guangzhou Uprising”. The main character of the incident, Sun Yat-sen, was born in Xiangshan County, Zhongshan, Guangdong on November 12, 1866. He and Huang Xing, who participated in the Huanghuagang Uprising, were comrades-in-arms, and Hu Hanmin was his assistant. Sun Yat-sen once signed the “Sun Wen Yuefei Declaration” on January 26, 1923, and the relevant old site is the Sun Yat-sen Memorial Hall, located in Canton.

The core of the research of this paper is the part of the Revolutionary History of the Republic of China facing "the Canton Canon". From the collected materials related to the history of the Canton Revolution, it unearthed knowledge classification, knowledge points, associations between knowledge points, knowledge attributes and knowledge attribute values, and finally established Knowledge Graph of Canton Revolutionary History in the Republic of China.

B. *Information Collection and Transformation of the Knowledge Graph of the Republic of China Canton Revolution*

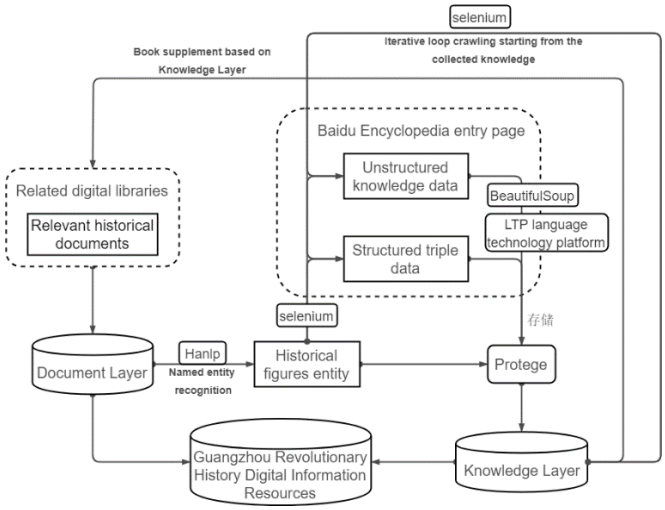

Fig. 4. Modeling flow chart of Canton Revolution history information of the Republic of China

The retrieval of the Canton Revolutionary Knowledge Graph in the Republic of China can be divided into two information layers: the document layer (historical document corpus) and the knowledge layer (knowledge graph data).

In terms of historical documents, we browsed and searched digital libraries such as "DUXIU Academic Search" with the core of "Republic of Canton Revolutionary History", based on the collected revolutionary events in Canton during the Republic of China, and core revolutionary figures who participated in the revolutionary events. Then, we use the collected revolutionary events in Canton during the Republic of China and the core revolutionary figures who participated in the revolutionary events as keywords to find documents that directly record the history of Canton's revolution, mainly biographies, Chinese history, Chinese Communist Party history, and a small part of literature and other disciplines, such as "The Biography of Sun Yat-sen", "Guangdong Revolutionary History Documents 1937-1940, CCP Canton Municipal Committee, Press and Periodical Materials Collection", "The History of the CCP in Guangzhou", "Guangzhou Before Dawn", etc. We downloaded and saved them and converted them into a relational database for Full-text retrieval through OCR text recognition and other techniques. As of the writing of this paper, there are 1241 historical documents in the Revolutionary History Document Corpus. As of the writing of this paper, there are 1241 historical documents in the Revolutionary History Document Corpus.

In terms of knowledge graph data information, NLP Tokenizer (NLP word segmentation) supported by the structured perceptron sequence labeling framework under the Chinese language processing package HanLp [11] is used to identify the named entity of each downloaded electronic document. The names of characters that may be related to the history of the Canton Revolution are obtained from the text of the books that record the historical events of the Canton Revolution. Based on the extracted characters, events, treaties and other knowledge entities, they are sequentially extracted based on the selenium library for multi-threaded Baidu Baike and Wikipedia web crawler knowledge structured triples. It is mainly divided into three parts: knowledge introduction, knowledge attributes, and knowledge relations for extraction.

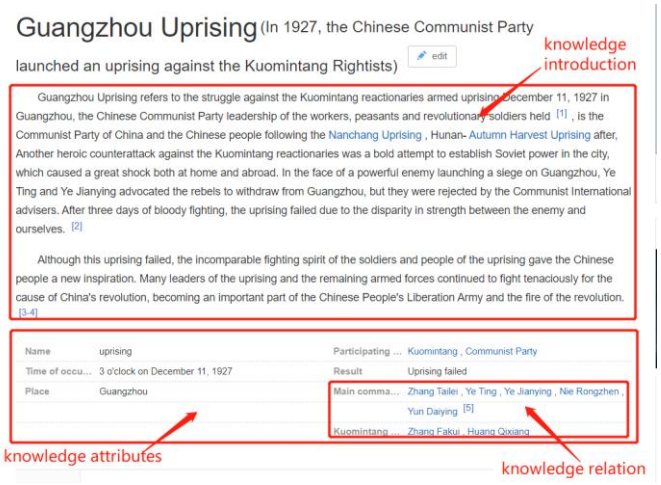

Fig. 5. "Guangzhou Uprising" encyclopedia entry crawler schematic diagram

Baidu Encyclopedia already has certain rules classification for the attributes of knowledge points and some relational information. Figure 5 is a schematic diagram of the Baidu Encyclopedia page of "Guangzhou Uprising". The HTML of the encyclopedia page is parsed to perform knowledge matching based on certain rules for each encyclopedia entry collected. As shown in Figure 5, the text above the attribute part is stored in the format of "Guangzhou Uprising, Knowledge Introduction, Guangzhou Uprising means..."; the knowledge attribute part can be extracted as "Guangzhou Uprising, Location, Guangzhou", "Guangzhou Uprising, Kuomintang Commander, Zhang Fakui, Huang Qixiang"; The knowledge relationship part can be extracted as "Guangzhou Uprising, related figures, Zhang Fakui", "Guangzhou Uprising, related figures, Huang Qixiang", within the agreed relationship range, knowledge part of the knowledge relationship can be extracted from the attributes to expand and obtain the exact information about events, people and other entities related to the history of the Guangzhou Revolution.

In addition, there is a large amount of unstructured knowledge in the knowledge base, which stores the life experiences of the characters and the entire process of events, which are used to intuitively show users the life deeds of the people related to the Canton Revolution and the full picture of the historical events of the Canton Revolution. Figure 6 is a schematic diagram of unstructured knowledge data related to the knowledge points of "Guangzhou Uprising". In order to

obtain a large number of such unstructured data, we used Python's BeautifulSoup package to parse the encyclopedia entry page of the corresponding knowledge point, and extracted the text from the part describing the life of the person and the event. Then the LTP language technology platform [12] developed by the Social Computing and Information Retrieval Research Center of Harbin Institute of Technology was used to recognize the time text of the character's life, and the segment where each time field is located is regarded as a knowledge node. By this method, multiple knowledge nodes are obtained and connected in series to obtain unstructured knowledge of the entire process of character's life and events.

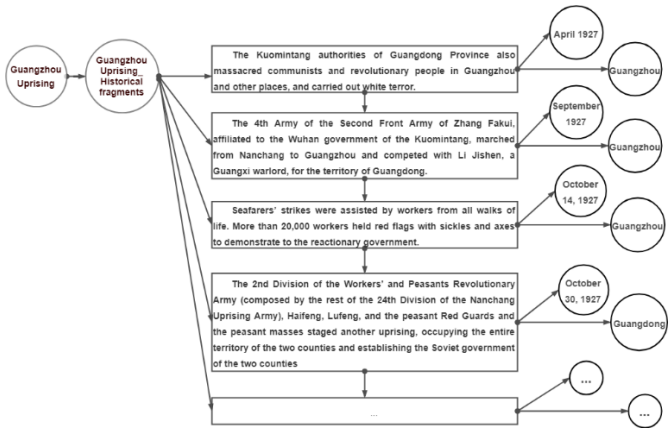

Fig. 6. Schematic diagram of unstructured knowledge data related to "Guangzhou Uprising"

Taking the collected historical events of the Canton Revolution and related figures as the starting point for crawling, crawled the documents and articles, literary works, cultural relics of the old site, archives and other information related to these events and figured on Baidu Encyclopedia to construct triple data to enrich the knowledge base of Canton revolutionary historical events. Finally, manual reading of relevant documents and consulting historians would supply some of the knowledge about Guangzhou Revolution history that the machine has missed. Then continued to iterate the above steps until the amount of knowledge in the knowledge base reaches the expected scale. As of this writing, the knowledge ontology has 8 super classes, 34 subclasses, 98 object property classes, 46 datatype property classes, 9,641 instances, and 16,435 ternary data items.

In terms of the knowledge base building specification, we used the ontology modeling tool Protégé [13] to build the ontology of the designed Guangzhou Revolutionary Knowledge Graph, and then called the Semantic Web Support Framework Jena [14] to parse the ontology OWL files generated by Protégé, and finally used SPARQL Protocol and RDF Query Language [15] to perform knowledge matching and extraction of the ontology data.

IV. Design of Double-Layer Retrieval Mode

Knowledge Graph has the advantages of semantic expression and analysis, which is especially suitable for revealing and retrieving the complex relationships among characters, events, places and times in the historical field of "the Canton Canon". At the same time, facing "the Canton Canon", the collection of the Canton Revolutionary Archives of the Republic of China can well supply the vacancy of the traditional knowledge graph platform in the user's implicit knowledge needs. Therefore, we constructed a double-layer retrieval mode to improve the relevance and diversity of retrieval knowledge. Through knowledge expansion, learning, and feedback of retrieval results, it can not only help users to adjust and revise their retrieval strategies, but also help users to mine more relevant tacit knowledge, and further improve the development and utilization rate of literature and books.

The Double-layer Retrieval Mode of document corpus based on Knowledge Graph is shown in Figure 7. On the basis of the information data at the Document Layer and the Knowledge Layer, we have realized the basic structure of the double-layer retrieval mode through the functional design of the full-text retrieval tool Lucene [16], the design of the knowledge matching and extraction process of the Knowledge Base. And here based on this, the N-triple document corpus retrieval is expanded. At the same time, we designed a semantic association retrieval process for some unstructured knowledge data that are easy to find in the retrieval process, and obtained a double-layer retrieval mode of document corpus based on the knowledge graph. This allows users to constantly switch back and forth between the Document Layer and the Knowledge Layer, and combine searches.

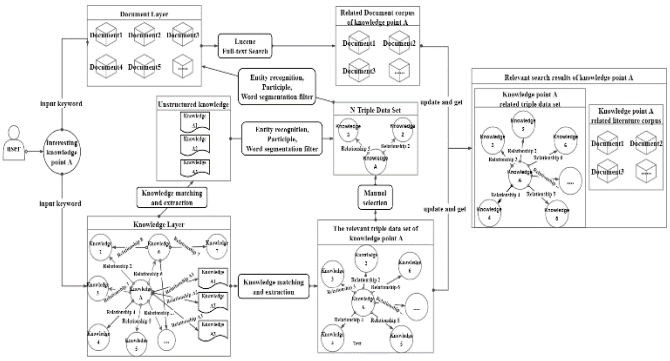

Fig. 7. Double-layer retrieval mode of document corpus based on Knowledge Graph

A. Overview of the Architecture of the Double-Layer Retrieval Mode

Figure 7 is the Double-layer retrieval mode of document corpus based on the Knowledge Graph. This mode has two information layers: a Knowledge Layer and a Document Layer, which respectively map the traditional keyword-based knowledge encyclopedia retrieval model and document retrieval model. Among them, knowledge retrieval is carried out on the Knowledge Layer. When the user passes in the knowledge point A that he wants to retrieve, A will perform knowledge matching and extraction in the knowledge base, and the user can get the relevant triple data set of the knowledge point A; The document retrieval is carried out at the document level. When the user passes in the knowledge point A that he wants to retrieve, it is matched in the document corpus based on the Lucene full-text search tool to obtain the relevant document corpus of the knowledge point A.

The bottom layer of the mode consists of two main information databases, a knowledge base and a document corpus, as well as subsidiary lexicons such as ontology knowledge lexicon, relational lexicon, and stop lexicon. The document corpus is a collection of document corpora in related fields. All the contents of the document corpus are stored in a relational database in byte stream format for standardization, which facilitates the subsequent full-text search to form a document corpus. Jena has strong analytic properties for ontology files. Consequently, knowledge points and knowledge relations are extracted through Jena's analysis of ontology files to construct ontology knowledge lexicon and relation lexicon to support subsequent word segmentation applications.

In the double-layer retrieval mode of document retrieval based on knowledge graph, when a certain knowledge point is searched, double-layer retrieval of knowledge layer and document layer is performed at the same time, and the sum of the obtained triple set and document corpus set is the final retrieval result. We use the document corpus and knowledge triple set as the retrieval result. This can not only encourage users to discover more relevant tacit knowledge through the high-precision matching segments of the literature corpus, but also supplement the lack of knowledge in user needs in the search results of traditional knowledge retrieval systems. Not only that, while breaking the limitations of traditional retrieval result knowledge triple data sets, it also improves the utilization rate of literature books.

B. *Knowledge Matching and Extraction Process in Knowledge Layer*

The process of knowledge matching and extraction is based on the RDF query language SPARQL and Jena TDB database, designing and coding SPARQL query statements which are used to search and filter the ontology data according to knowledge matching and extraction requirements, and generating ontology statement data. Finally, the ontology statement data is sorted and transmitted to the front end for display and data visualization.

Taking the "Guangzhou Uprising" knowledge point as an example, Figure 8 is a schematic diagram of the results of the knowledge extraction part of the "Guangzhou Uprising" example in knowledge layer. When the user passes in a knowledge point "Guangzhou Uprising", the SPARQL query sentence can be matched and sorted to get huge amounts of Relational triple like "Guangzhou Uprising, related-people, Ye Ting", "Guangzhou Uprising, relevant location, Guangzhou", "Guangzhou Uprising, start time, December 11, 1927" and some related attribute triple data such as "Guangzhou Uprising, result, The Uprising failed".

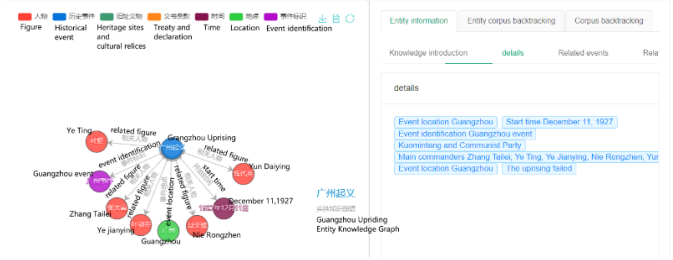

Fig. 8. Schematic diagram of the results of the knowledge extraction part of the "Guangzhou Uprising" example in Knowledge Layer

C. *Full-Text Retrieval Based on Lucene in Document Layer*

The Full-text retrieval based on Lucene is divided into two steps, namely the construction of the index of the document corpus and the Full-text retrieval of keyword segmentation. Figure 9 shows the specific flow chart of the Full-text search.

Before constructing the index, the sentence of the document needs to be segmented forward or backward, and the ontology knowledge base based on the Knowledge Graph is used as the segmentation rule for segmentation. The word segmentation component (Tokenizer) uses the IK-Analyzer tokenizer to segment all document content, and uses the ontology database to segment the tokens obtained. Then it passes the obtained word (Term) to the index component (Indexer) to create an index. The purpose of creating an index is to search, and ultimately it is necessary to search only the retrieved tokens. Index search refers to the user input keywords and sentences, and lexical analysis, grammatical analysis, and language processing of the query words, then search the index to obtain documents that conform to the syntax tree, and finally sort the results according to the relevance of the obtained documents and the keywords and sentences.

In the aspect of language processing, the ontology knowledge vocabulary and relation vocabulary of the knowledge graph are used for segmentation, and key entity data are extracted for retrieval. In terms of document ranking, the process of evaluating the importance of documents is the process of calculating term weights. The implementation of judging the relationship between the terms (Term) and obtaining the document relevance applies the Vector Space Model algorithm. The process of judging the relationship between the terms and obtaining the document relevance applies the Vector Space Model. Think of a document as a series of term, each term has a term weight. According to their weight in the document, different terms affect the scoring calculation of document relevance and arrange them according to the score, and the documents with high scores are displayed first.

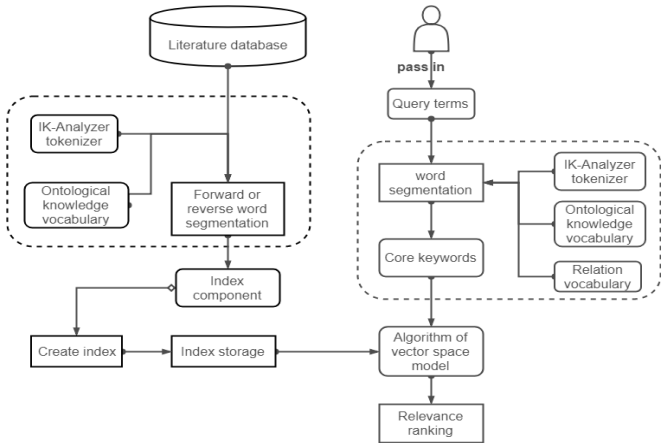

Fig. 9. Full-text retrieval flowchart

Taking "Guangzhou Uprising" as an example. Before constructing the index, using the IK tokenizer segment the document corpus forward or backward based on the ontology database to obtain tokens. The "Guangzhou Uprising" will be

passed as a token to Index component to create a lexical index of “Guangzhou Uprising”. When searching for “Guangzhou Uprising”, Lucene will find the index of the term, calculate the correlation coefficient of “Guangzhou Uprising” in each document, and obtain related documents related to “Guangzhou Uprising”, and press he correlation coefficients are arranged. And then the weight distribution of the word elements in the corpus is used for scoring, and the fragment scores of “Guangzhou Uprising” in each relevant document are obtained, and the document fragments are arranged according to the scores. Figure 10 is a schematic diagram of the Full-text search results of the “Guangzhou Uprising” document Layer. The scores of the first three documents are *Guangzhou Uprising Pictorial Collection* written by Wang Xiaoling, Jiang Bin, got score1=0.10908963, *Guangzhou Uprising* written by Huang Suisheng, got score2=0.09020603, *Looking at the Guangzhou Uprising* written by Xu Xiaolin, got score3= 0.07616874.

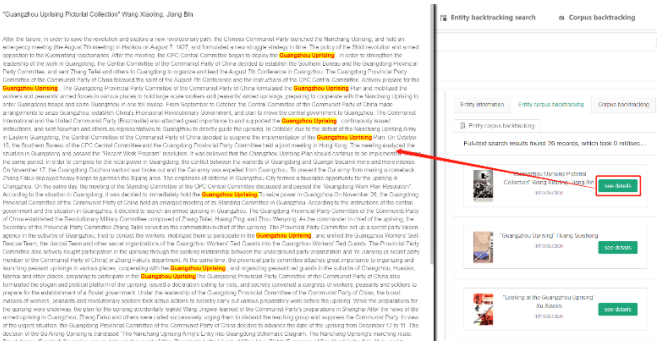

Fig. 10. Schematic diagram of the full text retrieval of the "Guangzhou Uprising" Document Layer

D. N-Triple Set Document Corpus Retrievar

The characteristic of knowledge graph is that it can express the relationship between entities most effectively and intuitively, connect a large amount of different kinds of information together to obtain a relationship network, and provide people with the ability to analyze problems from the perspective of “relationship”. The characteristic of knowledge graph is that it can express the relationship between entities most effectively and intuitively, connect a large amount of different kinds of information together to obtain a relationship network, and provide people with the ability to analyze problems from the perspective of “relationship”. Combining it with document retrieval, we expanded the N-triple set document corpus retrieval module in the two-layer retrieval mode based on the knowledge graph technology to take advantage of the knowledge relevance of the knowledge graph.

As shown in the double-layer retrieval mode in Figure 7, after the user has retrieved the knowledge point A of interest, he can directly select the knowledge set of interest in the relevant knowledge domain of the current knowledge point A. Then the system backend splits the points selected by the user into knowledge points and splices them into a long string before passing it into the document layer. And then use the Lucene full-text search tool to perform word segmentation and word segmentation relationship filtering based on the knowledge lexicon and the relational lexicon, retrieve the relevant document corpus set of the N-triple set and update it. Finally, search and construct a document corpus that is more relevant to the user’s demand domain according to user needs to obtain more knowledge information.

Take the “Guangzhou Uprising” knowledge points as an example. In the knowledge graph of the “Guangzhou Uprising”, the graph displays the knowledge collection of the N-triple set, and transfers the knowledge set to users. Then users can use the N-triple set of the “Guangzhou Uprising” knowledge graph to extract knowledge, and select “Guangzhou Uprising, related-people, Zhang Tailei”, “Guangzhou Uprising, related-people, Ye Jianying”, “Guangzhou Uprising, related-people, Ye Ting”, “Guangzhou Uprising, event location, Guangzhou”, “Guangzhou Uprising, start time, December 11, 1927” five triples. When these are input as the N-triple set for full-text search, the triple set will be spliced into the string “Guangzhou Uprising related-person Zhang Tailei related-person Ye Jianying related-person Ye Ting event location Guangzhou start time December 11, 1927”. After filtering by the relational lexicon, the string becomes “Guangzhou Uprising Zhang Tailei, Ye Jianying, Ye Ting, Guangzhou, December 11, 1927”. The word segmentation based on the knowledge vocabulary will segment the string into “Guangzhou Uprising”, “Zhang Tailei”, “Ye Jianying”, “Guangzhou”, “Ye Ting”, “December 11, 1927”, and then search through Lucene full-text technology a collection of literature corpora with higher scores associated with these six knowledge points is matched. Figure 11 is an example diagram of a simulated N-triple set document corpus retrieval model based on the knowledge ontology database of Guangzhou revolutionary historical events.

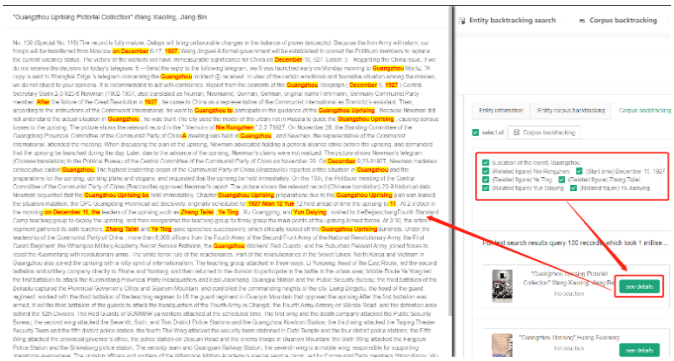

Fig. 11. The model diagram of the corpus retrieval model of the simulated N-triple set based on the knowledge ontology database of Guangzhou Revolutionary historical events

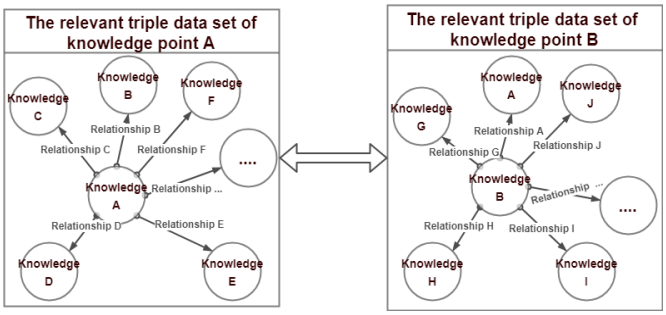

Fig. 12. Schematic diagram of Knowledge Graph jumping

At the same time, relying on the thinking-related advantages of the knowledge graph, users can continuously

jump between different semantically related knowledge graphs. Figure 12 is a schematic diagram of the jump of the knowledge graph. In the graph knowledge domain of knowledge point A, transfer the knowledge center from knowledge point A to knowledge B, and jump from the graph knowledge domain of knowledge point A to the graph knowledge domain of knowledge point B. Combining the thinking relevance characteristics of the knowledge graph with the retrieval of the N-triple set of document corpus to improve the knowledge relevance and diversity of the retrieval process. Through knowledge expansion, learning, and feedback of retrieval results to help users adjust and to correct retrieval strategies, and further supply the lack of knowledge in traditional knowledge retrieval system retrieval results on user needs.

In addition, some unstructured knowledge data can be obtained in the process of knowledge retrieval. This is a kind of information data between the Document Layer and the Knowledge Layer. In traditional digital libraries, such complex semantic information will not be further retrieved and analyzed. Generally, users extract the key of interest from it. After the word information is searched again or additional further information search is performed through other search engines such as Baidu Baike, which brings a lot of trouble to the knowledge search process. In this regard, we designed the extended search for such unstructured knowledge points in the retrieval mode, using the stop vocabulary of Harbin Institute of Technology as the stop vocabulary, and using the IK-Analyzer tokenizer to analyze the incoming unstructured vocabulary according to the knowledge entity vocabulary. After word segmentation and stop word filtering are performed on the structured knowledge data, it is constructed into the data form of the N-triple set, and then the Full-text search is performed.

V. Conclusion and Prospect

Based on the historical archives of Canton Revolution in the Republic of China, this article designs and constructs a knowledge graph of Canton Revolution history. Using the knowledge graph technology, the knowledge graph is associated with documentary corpus, which improves the relevance and diversity of knowledge in the retrieval process of traditional digital libraries. This can help users dig out more relevant tacit knowledge and supplement the missing knowledge in the retrieval results of traditional knowledge retrieval system in terms of users' demands; in addition, the transformation and combination of traditional digital library from literature information management to knowledge management greatly improves the development and utilization rate of literature books, which provides a very good solution for how to utilize the huge archival documents more efficiently and build a more open digital library. It opens up new ideas for the change from information management to knowledge management in China's digital library, which has certain theoretical and practical significance.

The "Digital Library of the Canton Canon", based on the knowledge graph, was designed and developed as a platform for the knowledge graph of "the Canton Canon", not only limited to the theme of the Canton Revolution of the Republic of China, but also as a demonstration application that can be gradually extended to any other theme contained in "the Canton Canon". After the completion of the second phase of "the Canton Canon", which is a compilation of the documents of the Republic of China, the research results of this project can be directly applied to provide a rich and effective knowledge service of "the Canton Canon" for experts and the public.

Acknowledgment

This work is supported by State Language Commission Informatization Project (No. YB135-123), Zhuhai Philosophy & Social Science Planning Project (No. 2019YB051), and Guangzhou Social Science Planning Project (No. 2019GZY26). It's also supported by the grants from the 13th five year plan project of philosophy and social science in Guangdong province (No. GD19CYY01), Provincial Key Platform and Major Scientific Research Project of Universities in Guangdong Province (Project No. 2020WTSCX122), and research program on the construction of an international education demonstration zone in the Guangdong-Hong Kong-Macao Greater Bay area (No. 2020WQYB030). It's supported by student's platform for innovation and entrepreneurship training program in Guangdong province (No. S201913177034) as well.

References

[1] The Canton Canon.[EB/OL][2020-10-10].http://gzdd.gzlib.gov.cn/.

[2] Zhu Qiang. Digital Library: The Prototype of the 21st Century Library-Introduction to the American "Digital Library Creation" Project [J]. Journal of University Libraries, 1995(4): 5.

[3] Sun Chengjian, Liu Gang. The beginning and development of China's digital library construction[J]. National Library of China, 2000(03):10-16.

[4] Ma Shuang. Discussion on the status quo and development of library digital resource integration under the network environment[J]. Information Records, 2020, 21(06): 41-42.

[5] Li Houqing, Kong Weitao. Research on the Evolution of Library Morphology and Its Development Direction[J]. University Library Work, 2020, 40(02): 1-6.

[6] Tu Jiaqi, Yang Xinya, Wang Yanli. Research on CNKI History and Development of CNKI[J]. Library Forum, 2019, 39(09):1-11.

[7] Singhal A. Introducing the knowledge graph: things, not strings[J]. Official google blog, 2012.

[8] Li Ying, Zhang Shuguang, Liu Yuxiu. The application of knowledge graphs in the analysis of discipline development[J]. Journal of Medical Postgraduates, 2013, 26(08): 875-877.

[9] Zhishi.me.[EB/OL][2020-10-10].http://zhishi.me/.

[10] Zhao Fang, Ren Ruijuan. System structure problems in the construction of digital library[J]. Information Magazine, 2004(10): 62-63.

[11] HanLP[EB/OL].[2020-10-10].http://www.hankcs.com/nlp/hanlp.html

[12] Wanxiang Che, Zhenghua Li, Ting Liu. LTP: A Chinese Language Technology Platform. In Proceedings of the Coling 2010:Demonstrations. 2010.08, pp13-16, Beijing, China.

[13] Stanford University School of Medicine. What is protégé.http://protege.stanford.edu/overview.[2012-12-01].

[14] Song Wei, Zhang Ming. A Concise Course of Semantic Web[M]. Shanghai: Higher Education Press, 2004.

[15] SPARQL[EB/OL].[2020-10-10].http://www.w3.org/TR/2008/REC-rdf-sparql-query-20080115/

[16] Lucene[EB/OL].[2020-10-10].http://lucene.apache.org/java/docs/index.html