Personalized Diagnostic Modal Discovery of Traditional Chinese Medicine Knowledge Graph

Yonghong Xie, Chang Yan, Dezheng Zhang
University of Science and Technology Beijing
Material domain knowledge engineering Beijing key laboratory
Beijing, China

Abstract—**Knowledge graph is a new research hotspot in the field of artificial intelligence. Traditional Chinese medicine (TCM) knowledge graph can well describe the relationship between symptoms, syndromes, etiology, treatment, prescriptions and so on. This paper proposes the storage structure of medical cases: four-tuple, path matrix, construct personalized knowledge graph of TCM, and on the basis of the knowledge graph of the basic theory of TCM, the personalized knowledge graph of famous TCM doctors is constructed, and the personalized diagnosis model of famous TCM doctors is found and verified by experiments. Finally, a data driven knowledge discovery method based on knowledge graph is proposed.**

Keywords-knowledge graph; personalized; storage structure; knowledge discovery

I. INTRODUCTION

Knowledge graph is a graphical representation of knowledge's internal structure and external relations. The purpose is to describe the entities and concepts that exist in the real world, as well as the relationships between these entities and concepts, and to capture and present the semantic relations between the concepts of the domain [1]. The concept of knowledge graph was formally proposed by Google in 2012. After 2013, the concept of knowledge graph was popularized in academia and industry, and played an important role in intelligent questions and answers, intelligence analysis, anti-fraud and so on, [2]. Yang Guoli [3], Qin Changjiang and Hou Hanqing [4] think that the knowledge graph is the combination of the theory and method of applied mathematics, graphics, information visualization, information science, etc., with the methods of citation analysis and co-occurrence analysis of metrology. It uses visual graph to visualization the multidisciplinary integration of the core structure of the subject, the history of development, the frontiers and the overall knowledge architecture. From the relevant literature, Hu Zewen [5] discusses the application of knowledge graph in the field of information science and its subfields, and briefly introduces the application of knowledge graph in other disciplines. Tang Jianmin [6], Zong Qianjin [7] and Xue Xiaofang [8] regard knowledge graph as a kind of research method, and think that scientific knowledge is the research object of knowledge graph. It combines the theory and method of mathematics and visualization synthetically, and combines with literature and scientific metrology to display the general graph, the relationship and evolution process of the subject research in the form of visual graph. It also can reveal the law of scientific development, grasp the trend of discipline development, and provide help for selecting research directions.

Traditional Chinese medicine (TCM) has a long history of thousands of years in China, the experience of old TCM doctor is the essence of TCM in the treasure. Gao Rui [9] thought that the TCM knowledge can be divided into: empirical knowledge assets, conceptual knowledge assets, systematic knowledge assets, practice knowledge assets. The customary knowledge assets are made up of implicit knowledge. These implicit knowledge are the practice in practice. Most of the experience of the famous TCM doctor is a kind of implicit knowledge. It is found out by the continuous social practice. It is very individualized. It is difficult to summarize and refine it, if this part of implicit knowledge is explicit, it will play a great role in knowledge inheritance. The establishment of the "12th Five-Year" study on the inheritance study of academic thoughts and academic thoughts of TCM doctors "by the Ministry of science and technology of the state is a promotion the informatization and digitization of the Chinese medicine field. Due to the complexity of the relationships between TCM knowledge and the limitations of traditional storage structure, it is not conducive to the application of knowledge. The knowledge graph, that can solve this problem well, pay attention to knowledge graph relationship between the entities, entity extraction and knowledge discovery technology can quickly construct large-scale, high quality knowledge graph [2]. Knowledge graph can be enriched and expanded on the basis of ontology, which highlights and emphasizes the relationship between concepts and concepts, and the knowledge graph is based on the ontology, adding more abundant information about the entity [10]. The specific implementation of knowledge graph can be realized by means of a graph database. In this paper, we use the Neo4j database.

A. Knowledge Graph of Basic Theory of TCM

The knowledge graph of basic theory of TCM is established on the basis of TCM Ontology. The atlas is guided by the theory of yin and Yang and five elements. It is centered on the five organs. It includes four parts of [10], the cognitive method of TCM, the physiology of TCM, the pathology of TCM and syndrome differentiation and treatment. The basic structure of TCM basic theory knowledge graph consists of the concept hierarchy diagram GM and entity relationship diagram GE, namely KG=<GM, GE>. Among them, the concept hierarchy diagram represents the hierarchical structure of TCM ontology

This work is supported by the Beijing Key Laboratory of Knowledge Engineering for Materials Science under Grant 2017YFB1002304

978-1-5386-8097-1/18/$31.00 ©2018 IEEE

Authorized licensed use limited to: SOUTH CHINA UNIVERSITY OF TECHNOLOGY. Downloaded on May 13,2023 at 03:21:21 UTC from IEEE Xplore. Restrictions apply.

concept; entity relation diagram represents the relationship between TCM entities and their relationships.

B. Neo4j Graph Database

Neo4j is a high performance NOSQL graphics database. It is an embedded, disk based Java persistence engine with complete transactional features, but it stores structured data on a network (called a graph from a mathematical point of view) instead of a table. The performance of traversal without the impact of the size of the graph data is an important feature of Neo4j, which makes Neo4j an ideal database for solving graph problems, even if the data set is very large.

Neo4j can be used in various fields such as social computing, recommendation engines, telecommunications, authorization and access control, routing and logistics, product catalogues, data center management, career management, fraud detection, public security and geographic space, which have proved to be an ideal choice for dealing with complex data. At present, Neo4j occupies the position of NO.1 in the whole graph database.

II. ACCESS TO MEDICAL DATA AND NORMALIZATION

Medical cases are the data records of the practice of knowledge application in TCM. It contains the results of solving practical problems in the application of TCM. We try to find the pattern of treatment according to syndrome differentiation of TCM by combining the medical cases with the knowledge graph of the TCM. Through the national 10th and 11th Five-Year plan, we have a large number of high quality medical records of TCM doctors. First, we standardize the symptoms, syndromes, drugs and other information of each medical case according to the language standard of the TCM knowledge graph, and store the standard medical cases.

A. Standardized medical case

According to the description of each symptom, syndrome, treatment method and TCM in the medical case, and combining the method of semantic similarity calculation, the corresponding terms in the basic knowledge graph of Chinese medicine determine its standardized expression and further standardize the medical case words into a set of standard words.

For example, a medical case is normalized. For example, symptoms "stomach fullness, dull pain, recurrent attacks, exacerbation in the past two weeks." Pain in hunger and postprandial, dull pain, occasionally tingling. The tongue is dark red, yellow and white tongue coat, and the roots are greasy and the veins are deep. It is standardized as "stomachache, belch, dry mouth, swallowing bitterness, gastric anorexia, and dry", and further dismantling as a collection of words {stomach pain, belch, dry mouth, pharynx bitterness, gastric anorexia, and dry stool}; syndrome " interresistance of phlegm and stasis, unbalance of stomach " standardized as "stasis and stomach collateral syndrome", dismantling as a collection of words {syndrome of static blood in stomach collaterals} etc.

B. Standardized medical case storage structure

The standardized medical records are stored in accordance with the structure shown in Table I.

TABLE I. STANDARDIZATION OF MEDICAL RECORDS STORAGE STRUCTURE

Table Name	**Symptom in Medical Cases / Syndrome in Medical Cases / Treatment in Medical Cases / Name of Chinese Medicine in Medical Cases**	
Data Storage	***Basic theory database of TCM***	
Column name	Data Type	Remarks
ID	varchar	ID of medical cases
Name	varchar	Symptom / Syndrome /Treatment/ Chinese Medicine

Normalization of all medical records of an old Chinese medicine is shown in Table II, III, IV, V.

TABLE II. PARTIAL THE COMPLETE DATA BASIS SYMPTOMS

ID	**Name**
5498	palpitation
5498	feverish palms and soles
5498	dry mouth
5498	dry stool
5498	dreaminess
5498	red tongue
5498	yellow and greasy tongue coating
5498	thready pulse
5498	five upset hot
5498	dry lips
5498	dry mouth and throat
5498	thirst
5498	constipation
5504	dizziness
5504	headache
…	…

TABLE III. PARTIAL THE COMPLETE DATA BASIS SYNDROME

ID	**Name**
5498	syndrome of endogenous heat due to yin deficiency
5498	syndrome of deficiency-heat due to the heart-yin
5504	syndrome of yin deficiency of liver and kidney
…	…

TABLE IV. PARTIAL THE COMPLETE DATA BASIS MEDICAL TREATMENT

ID	**Name**
5498	nourishing yin and clearing heat
5498	soothe the nerves and diazepam palpitation
5498	tranquilizing and sedating the Mind
5504	calm the liver and suppress yang
…	…

TABLE V. PARTIAL THE COMPLETE DATA BASIS CHINESE MEDICINE

ID	**Name**
5498	figwort
5498	salvia miltiorrhiza
5498	kuh-seng
5498	yuanhu
5498	scutellaria baicalensis
5498	coptis chinensis
5498	radix glehniae
5498	codonopsis pilosula
5498	radix ophiopogonis
5498	schizandra sinensis Baill
5498	aceranthus sagittatus S. et Z
…	…

Authorized licensed use limited to: SOUTH CHINA UNIVERSITY OF TECHNOLOGY. Downloaded on May 13,2023 at 03:21:21 UTC from IEEE Xplore. Restrictions apply.

III. OBTAIN THE MEDICAL CASE NODE AND GET THE PATH AND STORAGE

Each normalized medical case was put into knowledge graph of Chinese medicine for path query.

A. Path Query

A path query of a TCM doctor's medical cases will be carried out in the following way:

1) With symptom as the starting node, syndrome as the termination node, the path between the starting node and the terminating node is querying in the knowledge graph of TCM.

2) With syndrome as the starting node, treatment as the termination node, the path between the starting node and the terminating node is querying in the knowledge graph of TCM.

3) With treatment as the starting point and Chinese medicine as the termination node, the path between the starting node and the terminating node is querying in the knowledge graph of TCM.

B. Storage of Path Query Results

The path results of each case are stored in the form of path matrix and path four-tuple.

The path matrix is used to store path information between nodes. The basic structure of the path matrix is composed of node ID and path number. The node ID includes the starting node ID and the termination node ID.

The path matrix is symmetric, the diagonal element is 0, and the path matrix is shown in Fig. 1.

$$MatrixPath = \begin{bmatrix} 0 & ID1 & ID2 & ID3 & \cdots & IDn \\ ID1 & 0 & m1 & m2 & \cdots & m4 \\ ID2 & m1 & 0 & m3 & \cdots & m5 \\ ID3 & m2 & m3 & 0 & \cdots & m6 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ IDn & m4 & m5 & m6 & \cdots & 0 \end{bmatrix}$$

Figure 1. Knowledge graph of TCM(Partial data)

Among them, MatrixPath represents the path matrix; IDi represents the ID of the NO. i node, 1≤i≤n, n represents the number of nodes; m1, m2, m3, m4, m5, m6... represents the number of path between the corresponding nodes.

The path four-tuple is used to store the middle information of the path. The path four-tuple is expressed as:

$$TuplePath = \{<S>、 <E>、<T>、<V>\} \quad (1)$$

Among them, <S> represents the ID set of starting nodes, <E> represents the ID set of terminating nodes, <T> represents the path type, <V> represents the information of intermediate nodes in the path, V= {(M, C)}, M represents the ID of the intermediate node in the path, and C represents the frequency that the ID of the intermediate node appears in the path.

TABLE VI. PARTIAL DATA OF A MEDICAL SYMPTOM-SYNDROME PATH QUERY RESULT MATRIX

0	**1771**	**3072**	**3595**	**21960**	**1202**	**837**	**886**	**1398**	**22524**
1771	0	2	1	2	0	0	0	0	0
3072	2	0	0	0	1	1	1	2	2
3595	1	0	0	0	1	1	1	1	0
21960	2	0	0	0	1	1	1	2	2
1202	0	1	1	1	0	0	0	0	0
837	0	1	1	1	0	0	0	0	0
886	0	1	1	1	0	0	0	0	0
1398	0	2	1	2	0	0	0	0	0
22524	0	2	0	2	0	0	0	0	0
…	…	…	…	…	…	…	…	…	…

TABLE VII. FOUR-TUPLE OF A MEDICAL SYMPTOM-SYNDROME PATH QUERY RESULT MATRIX

Starting Node	**Terminal Node**	**Type**	**Value**	
1771	3072	Symptoms-disease's nature, Syndrome- disease's nature	21542,1	21589,1
1771	3595	Syndromes-symptoms, Subconcept	3591,1	
1771	21960	Symptoms-disease's nature, Syndrome- disease's nature	21542,1	21589,1
1202	3595	Syndromes-symptoms, Subconcept	3591,1	
1202	3072	Symptoms-disease's nature, Syndrome- disease's nature	21542,1	
1202	21960	Symptoms-disease's nature, Syndrome- disease's nature	21542,1	
837	3595	Syndromes-symptoms, Subconcept	3591,1	
837	3072	Symptoms-disease's nature, Syndrome- disease's nature	21542,1	
837	21960	Symptoms-disease's nature, Syndrome- disease's nature	21542,1	
886	3595	Syndromes-symptoms, Subconcept	3591,1	
886	3072	Symptoms-disease's nature, Syndrome- disease's nature	21542,1	
886	21960	Symptoms-disease's nature, Syndrome- disease's nature	21542,1	
1398	3595	Syndromes-symptoms, Subconcept	3591,1	
1398	3072	Symptoms-disease's nature, Syndrome- disease's nature	21542,1	21589,1
1398	21960	Symptoms-disease's nature, Syndrome- disease's nature	21542,1	21589,1
22524	3072	Symptoms-disease's nature, Syndrome- disease's nature	21542,1	21589,1
22524	21960	Symptoms-disease's nature, Syndrome- disease's nature	21542,1	21589,1
22524	21960	Symptoms-disease's nature, Syndrome- disease's nature	21542,1	21589,1
…	…	…	…	…

Store the search results of each medical record path of an old Chinese medicine doctor. Each medical case receives a set

Authorized licensed use limited to: SOUTH CHINA UNIVERSITY OF TECHNOLOGY. Downloaded on May 13,2023 at 03:21:21 UTC from IEEE Xplore. Restrictions apply.

of data, including symptom syndrome matrix and four-tuple, syndrome - treatment matrix and four-tuple, treatment - Chinese medicine matrix and four-tuple. After processing 102 medical records of an old Chinese doctor, a total of 102 sets of data were obtained. As shown in Table VI, Table VII. For example, there are two paths between No. 1771 and No. 3072.

The four-tuple describes the intermediate node of the path. It is possible that there are multiple paths of different intermediate nodes in the two nodes. For example, the same path type exists between No. 1771 and No. 3072, and one path passes through No. 21542 and another passes through No. 21589.

IV. Overlay All Medical Path Results

After all the medical cases completed the path query and stored, the matrices and four-tuple of each medical case were superimposed to obtain the general matrix and the general four-tuple of all medical records.

102 sets of data from path query results of a TCM doctor were superimposed to obtain the symptom-syndrome general matrix and general four-tuple, syndrome-treatment general matrix and general four-tuple, and treatment-Chinese medicine general matrix and general four-tuple. As shown in Table VIII, Table IX.

V. Using the Template

Build a personalized knowledge graph. The build process is as follows:

a) Traversing the matrix, we get the row node ID and column node ID whose number of paths is not 0. Since the matrix is a symmetrical matrix, we can only traverse half of the matrix.

b) The row node ID and column node ID are used as query conditions to find corresponding entities in TCM knowledge graph;

c) In the search between two entities, the naming format of the relationship is "doctorName_relationName", the relation type is declared in the relational type table, the side direction is determined by the direction of the relationship, and the corresponding matrix element is the weight of the edge.

A. Traversing General Path Matrix

Traversing general path matrix includes:

1) Traversing the syndrome-symptom general matrix to obtain the row ID and column ID of a node, then find the corresponding entity in the TCM knowledge graph, and construct an edge between the two entities whose relationship type is "ckj_syndrome-symptom"(ckj is the abbreviation of the doctor's name). The relationship type is stated in the relationship type table. The direction of the relationship is based on the direction of the "syndrome-symptom" and the corresponding element in symptom-sense matrix is used as the edge weight.

TABLE VIII. Partial Data of All Medical Records Superposed Symptom-Syndrome Matrix

0	**1771**	**3072**	**3595**	**21960**	**1202**	**837**	**886**	**1398**	**22524**
1771	0	6	3	6	0	0	0	0	0
3072	6	0	0	0	2	1	3	4	4
3595	3	0	0	0	2	1	3	2	0
21960	6	0	0	0	2	1	2	2	2
1202	0	2	2	2	0	0	0	0	0
837	0	1	1	1	0	0	0	0	0
886	0	3	3	2	0	0	0	0	0
1398	0	4	2	2	0	0	0	0	0
22524	0	4	0	2	0	0	0	0	0
…	…	…	…	…	…	…	…	…	…

TABLE IX. Partial Four Tuple of All Medical Records Superposed Symptom-Syndrome Matrix

Starting Node	**Terminal Node**	**Type**	**Value**	
1771	3072	Symptoms-disease's nature, Syndrome- disease's nature	21542,3	21589,3
1771	3595	Syndromes-symptoms, Subconcept	3591,3	
1771	21960	Symptoms-disease's nature, Syndrome- disease's nature	21542,3	21589,3
1202	3595	Syndromes-symptoms, Subconcept	3591,2	
1202	3072	Symptoms-disease's nature, Syndrome- disease's nature	21542,2	
1202	21960	Symptoms-disease's nature, Syndrome- disease's nature	21542,2	
837	3595	Syndromes-symptoms, Subconcept	3591,1	
837	3072	Symptoms-disease's nature, Syndrome- disease's nature	21542,1	
837	21960	Symptoms-disease's nature, Syndrome- disease's nature	21542,1	
886	3595	Syndromes-symptoms, Subconcept	3591,3	
886	3072	Symptoms-disease's nature, Syndrome- disease's nature	21542,3	
886	21960	Symptoms-disease's nature, Syndrome- disease's nature	21542,2	
1398	3595	Syndromes-symptoms, Subconcept	3591,2	
1398	3072	Symptoms-disease's nature, Syndrome- disease's nature	21542,2	21589,2
1398	21960	Symptoms-disease's nature, Syndrome- disease's nature	21542,1	21589,1
22524	3072	Symptoms-disease's nature, Syndrome- disease's nature	21542,2	21589,2
22524	21960	Symptoms-disease's nature, Syndrome- disease's nature	21542,1	21589,1
…	…	…	…	…

2) Traversing the syndrome-treatment general matrix to obtain the row ID and column ID of a node, and then find the corresponding entity in the TCM knowledge graph, and constructing the edge of the relationship between the two

Authorized licensed use limited to: SOUTH CHINA UNIVERSITY OF TECHNOLOGY. Downloaded on May 13,2023 at 03:21:21 UTC from IEEE Xplore. Restrictions apply.

entities whose relationship type is "ckj_syndrome-treatment", the relationship type is stated in the relationship in the type table, The direction of the relationship is based on the direction of the "syndrome-treatment" and the corresponding element in syndrome-treatment matrix is used as the edge weight.

3) Traversing the treatment-chinese_medicine general matrix to obtain the row ID and column ID of a node, and then find the corresponding entity in the TCM knowledge graph, and constructing the edge of the relationship between the two entities whose relationship type is "ckj_treatment-chinese_medicine", the relationship type is stated in the relationship in the type table, The direction of the relationship is based on the direction of the "treatment-chinese_medicine" and the corresponding element in syndrome-treatment matrix is used as the edge weight.

B. Using Neo4j Graphics Database to Build Personalized Knowledge Graph

The construction of personalized knowledge graph of a TCM doctor was completed in the Neo4j graphic database. As shown in Fig. 2 and Fig. 3.

VI. RESULT ANALYSIS

In this experiment, we can get the following analysis results and conclusions.

A. Analysis of the Main Symptoms

According to TCM dialectics theory, symptoms are the main basis for TCM identify specific syndromes. Although the symptoms and signs of the syndrome and disease are varied, but for a specific syndrome, the symptoms can be divided into two major categories. The main symptom refers to the main symptoms and signs of the disease, reflecting the main contradictions of the disease and having a very close and direct connection with the nature of the disease. Therefore, in the dialectical process of TCM, the main disease has a major position in many clinical manifestations and, to a certain extent, plays a decisive role in other symptoms and signs.

A medical case has multiple symptoms. The path query of TCM knowledge graph shows that there may be no path between a specific symptom and syndrome. When conducting a path query, a group of symptoms were entered, and only a few symptoms and syndromes can be connected, and there was no path between most symptoms and syndromes. It can be seen that in TCM knowledge graph, only a small part of the symptoms, which is included in the main symptoms, play a role in reasoning.

Figure 2. Personalized subgraph

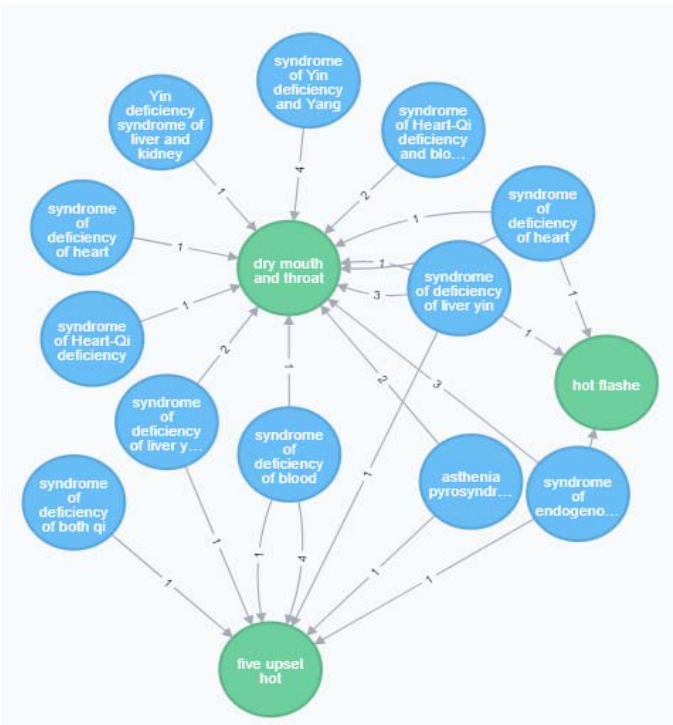

Figure 3. Personalized weighted edge subgraph

Authorized licensed use limited to: SOUTH CHINA UNIVERSITY OF TECHNOLOGY. Downloaded on May 13,2023 at 03:21:21 UTC from IEEE Xplore. Restrictions apply.

There are 1933 symptoms and 506 main symptoms in the medical case of a TCM doctor. There are only 1 to 11 main symptoms in each medical case, and the majority of the medical cases only have 1 to 5 main symptoms. Since the symptom charts of the 102 medical records were too large, 12 data were selected as representatives, as shown in Fig. 4.

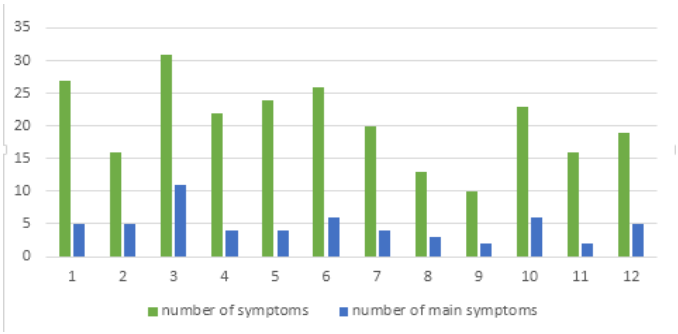

Figure 4. Histogram of symptoms and main symptoms

A further analysis of 102 medical records revealed that 24.7% of the main symptoms were tongue image. In TCM cases, tongue accounts for only 6.5% of all symptoms. So in these 102 cases, tongue is the main component of the main symptoms. From the above analysis, we can see that a TCM doctor pays attention to the diagnosis of tongue. This conclusion is consistent with the conclusion drawn from the study of the TCM heritage team.

B. *Analysis of Intermediate Node*

Because there are too many intermediate nodes for symptoms and syndromes, we mainly analyze the intermediate nodes of syndromes and treatment. Analyze the general four-tuple of syndrome-treatment for statistics. According to Fig. 5, a TCM doctor for who is good at the cardiovascular disease. The main means of diagnosis and treatment for the cardiovascular disease are promoting blood circulation for removing blood stasis (NO. 3793), invigorate the circulation of blood and activate stagnation (NO. 3806), invigorate the circulation of blood and relieving pain (NO. 3812).

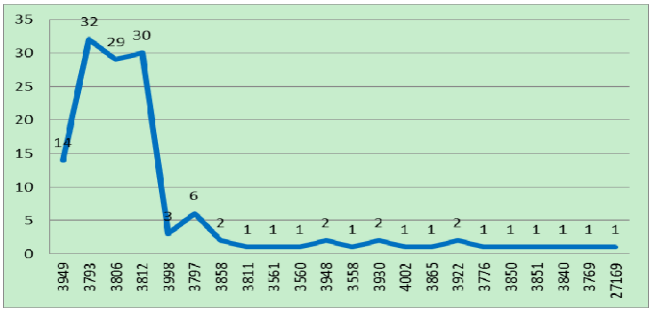

Figure 5. The intermediate node line graph between syndrome and therapy

From the above analysis, we can see that a TCM doctor has adopted different methods in the treatment of the disease. This conclusion is consistent with the conclusion drawn from the study of the TCM heritage team.

C. *Diagnostic Mode Verification*

According to the syndrome analysis of medical cases, it can be known that a TCM doctor is good at heart pulse obstruction syndrome, so three medical cases of other doctors with the syndromes of heart pulse obstruction are tested and compared with TCM dialectic method. The main input is the symptoms and symptoms of the medical cases.

1) The NO. 1 medical case and the test result are shown in Table X and Table XI.

TABLE X. THE NO. 1 MEDICAL CASE

Symptom	Syndrome
dizziness, headache, chest tightness, shortness of breath, weakness, swelling pain of head-eye, numbness, edema, dry mouth, bitter taste, polyuria at night, constipation	Syndrome of blood stasis and heart pulse、Syndrome of obstruction of heart and veins

TABLE XI. THE REASONING RESULT OF NO. 1 MEDICAL CASE

Traditional Dialectical Methods of TCM		Individualized Diagnosis Model of a TCM Doctor	
Syndrome	***count***	***Syndrome***	***count***
syndrome of deficiency of heart blood	6	syndrome of blockade of heart vessel	59
syndrome of deficiency of heart-YANG	6	syndrome of heart qi Yin deficiency and blood stasis	26
syndrome of deficiency of blood	4	syndrome of deficiency of heart-YANG	14
syndrome of heatstroke	4	syndrome of heart qi Yin deficiency	13
syndrome of phlegm-heat obstructing lung	4	syndrome of deficiency of heart qi and blood stasis	11
syndrome of deficiency of heart qi	4	syndrome of deficiency of heart qi	8

2) The NO. 2 medical case and the test result are shown in Table XII and Table XIII.

TABLE XII. THE NO. 2 MEDICAL CASE

Symptom	Syndrome
dizziness, numbness, fatigued spirit and lack of strength, dry mouth, chest tightness, bad sleep	syndrome of blockade of heart vessel

TABLE XIII. THE REASONING RESULT OF NO. 2 MEDICAL CASE

Traditional Dialectical Methods of TCM		Individualized Diagnosis Model of a TCM Doctor	
Syndrome	***count***	***Syndrome***	***count***
syndrome of qi deficiency of heart and lung	4	syndrome of blockade of heart vessel	38
syndrome of phlegm-heat obstructing lung	3	syndrome of deficiency of spleen qi	21
syndrome of exuberance of liver fire	3	syndrome of deficiency of heart-YANG	13
syndrome of yang deficiency of heart and kidney	3	syndrome of heart qi Yin deficiency	10
syndrome of deficiency of heart-YANG	3	syndrome of stagnation of liver qi and spleen deficiency	9
syndrome of blockade of heart vessel	3	syndrome of endogenous heat due to yin deficiency	6

Authorized licensed use limited to: SOUTH CHINA UNIVERSITY OF TECHNOLOGY. Downloaded on May 13,2023 at 03:21:21 UTC from IEEE Xplore. Restrictions apply.

3) The NO. 3 medical case and the test result are shown in Table XIV and Table XV.

TABLE XIV. THE NO. 3 MEDICAL CASE

Symptom	Syndrome
chest pain, chest tightness, shortness of breath, palpitation, weakness, chills	Syndrome of heart and vein of cold arthralgia、syndrome of blockade of heart vessel

TABLE XV. THE REASONING RESULT OF NO. 3 MEDICAL CASE

Syndrome-Differentiation by Eight Principles		Seven Steps of Syndrome-Differentiation		Path Pruning		Individualized Diagnosis Model of a TCM Doctor	
Syndrome	*count*	*Syndrome*	*count*	*Syndrome*	*count*	*Syndrome*	*count*
syndrome of yang deficiency of heart and kidney	3	syndrome of qi deficiency of heart and lung	3	syndrome of blockade of heart vessel	6	syndrome of blockade of heart vessel	38
syndrome of qi deficiency of heart and lung	3	syndrome of deficiency of heart qi and blood stasis	3	syndrome of yang deficiency of heart and kidney	6	syndrome of deficiency of heart qi and blood stasis	17
syndrome of deficiency of heart-YANG	3	syndrome of deficiency of heart-YANG	3	syndrome of deficiency of heart-YANG	6	syndrome of deficiency of heart-YANG	12
syndrome of blockade of heart vessel	3	syndrome of blockade of heart vessel	3	syndrome of deficiency of heart blood	2	syndrome of stagnation of liver qi and spleen deficiency	8
syndrome of water-rheum collecting internally	2	syndrome of deficiency of heart qi	2	syndrome of malnutrition of heart spirit	2	syndrome of heart qi yin deficiency	7

Through the verification of the above three medical cases, it is known that the knowledge discovery method driven by medical cases data which is based on the TCM knowledge graph, can effectively preserve the diagnosis model of the old Chinese medicine, and superimpose the diagnostic model on the TCM theoretical knowledge atlas more effectively for the reasoning and diagnosis of medical cases.

From the above results, we can find that the discovery method of personalized diagnosis model of TCM knowledge atlas, driven by medical cases, can find idea of the TCM treatment according to syndrome differentiation, and can also find the diagnostic model that the TCM heritage team has not found.

Starting from the theoretical knowledge of TCM, based on the characteristics of the TCM knowledge graph, driven by the TCM medical case, found the TCM diagnosis model. It has opened up a new idea for the research of the TCM personalized diagnosis model in the field of TCM, and it also applies to other fields based on knowledge graph, data-driven personalized knowledge discovery, and promotes the pace of modern information construction in TCM.

D. Data-Driven Personalized Knowledge Discovery Method for Knowledge Graph

When the knowledge graph is applied in the field of TCM, the method and experimental process are found on the basis of the TCM medical case, based on the personalized diagnosis model of TCM. This section further studies this method, and proposes a data driven knowledge mapping method for personalized knowledge discovery.

The module structure and process of the knowledge discovery method are as follows:

1) Get standardized module

It includes data acquisition module and data specification module. The data acquisition module is used to obtain at least one portion of the data; the data specification module is used to standardize the data obtained by querying the corresponding entities in the knowledge graph.

2) Get node module

It is used to standardize data, combine normalized data with practical application scenarios, and select normalized data as starting node and terminating node.

3) Get path module

According to the knowledge graph, the path between the starting node and the termination node is obtained as piece of data.

4) Mode determination module

It is used to determine the personalized mode based on each piece of data obtained.

The mode determination module includes:

a) A path storage unit for storing the path of each acquired data as a path matrix and a path four-tuple. The path matrix is used to store the path information between nodes. The basic structure of the path matrix is composed of node ID and path number. The node ID includes the starting node ID and the termination node ID. (See Fig. 1 for details)

b) The superposed path unit is used to superimpose the path matrix of each data stored to get the general path matrix of all data, and the general path four-tuples of each piece of data are superimposed to get the total four-tuples of the path of all data.

c) The pattern determination unit is used to determine the personalized knowledge pattern based on the general path matrix and the general four tuples.The mode determination unit includes: Traversing subunits, used to traverse the path matrix, to obtain row node IDs and column node IDs of the elements whose number of paths is not 0; Finding subunits, using the obtained row node IDs and column node IDs as query conditions, find the entities corresponding to row node IDs and column node IDs in the TCM knowledge graph; A subunit is built to construct a personalized knowledge relationship between the found the entity corresponding to the row node ID and the column node ID, based on the starting node ID in the path total four-tuples, the terminating node ID, the path type, and the middle node information in the path. And the number of corresponding paths in the total matrix of

Authorized licensed use limited to: SOUTH CHINA UNIVERSITY OF TECHNOLOGY. Downloaded on May 13,2023 at 03:21:21 UTC from IEEE Xplore. Restrictions apply.

the path is used as the weight of the edge of the personalized knowledge relationship.

From the above modules and processes, the difference between the knowledge discovery method and the other methods is not to find the implicit knowledge and the new knowledge by using data mining in the knowledge base, but to query in the knowledge graph with the external data. Based on the query results, we build a personalized knowledge graph based on the original knowledge graph and discover new knowledge.

The flow of this method is shown in Fig. 6.

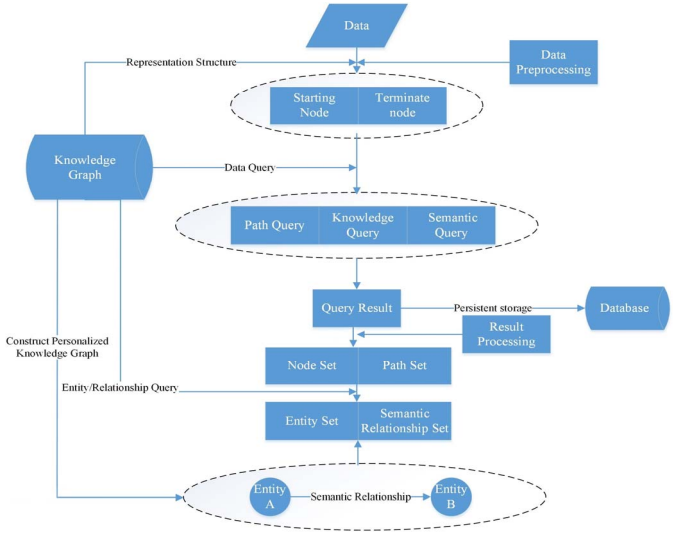

Figure 6. Personalized knowledge discovery of data driven knowledge graph flow chart

VII. CONCLUSION

Based on the TCM knowledge graph, this article studies the discovery techniques of personalized knowledge graph of TCM, driven by TCM medical cases, and has carried out an experimental verification. This method is beneficial to automatically discover the pattern of treatment according to syndrome differentiation of TCM under the background of big data. It opens up a path for the effective inheritance of the experience of TCM doctors. In the future, we will further modify the representation method of knowledge graph to better discover and express the experience knowledge of TCM doctors.

REFERENCES

[1] Zhang L. Knowledge graph theory and structural parsing. University of Twente, 2002.

[2] Qi GuiLin, Gao Huan, Wu TianXing. The Research Advances of Knowledge Graph. technology intelligence engineering, 2017, 3(1):4-25.

[3] Yang Guoli, Li Pin, Liu Jing. Mapping Knowledge Domain——A New Field of Scientometrics. science popularization, 2010, 5(4):28-34.

[4] Qin Changjiang, Hou Hanqing. Knowledge Map——A New Field of Information Management and Knowledge Management. Journal of Academic Libraries, 2009, 27(1):30-37.

[5] Hu Zewen, Sun Jianjun, Wu Yishan. A Survey of the Application Research of Domestic Knowledge Mapping. Library and Information Service, 2013, 57(3):131-137.

[6] Tang Jianmin. Mapping of Subject Knowledge Atlas and Application in Subject Development Monitoring and Evaluation. Information Studies:Theory & Application, 2009, 32(10):55-59.

[7] Zong Qianjin, Yuan Qinjian, Shen Hongzhou, Shu Xiaoyun. Review of Information Science Research in China from 2001 to 2010 and Prospects for Future Research: A Study of Contemporary Trends in the Knowledge Mapping-based Discipline. Information and Documentation Services, 2012, 33(1):10-15.

[8] Xue Xiaofang. Theory, Methods and Tools of Knowledge Visualization and Application Study in Military Medicine. China Academy of Military Medical Sciences, 2014.

[9] Gao Rui, Xu Yongmei, Zhu Zhengxiang, Gu Jifa. Exploring Methods to Summarize and Inherit Experience and Thoughts of Elder and Famous Doctors in Traditional Chinese Medicine. Chinese Journal of Experimental Traditional Medical Formulae,2011,17(08):275-278.

[10] Zhang DeZheng, Xie Yonghong, Li Man, Shi Chuan. Construction of Knowledge Graph of Traditional Chinese Medicine Based on the Ontology. Technology Intelligence Engineering, 2017, 3(1):35-42.

Authorized licensed use limited to: SOUTH CHINA UNIVERSITY OF TECHNOLOGY. Downloaded on May 13,2023 at 03:21:21 UTC from IEEE Xplore. Restrictions apply.