# Knowledge Mapping of Medicinal Plants Based on Artificial Neural Network

**Lei Miao**

*Office of Academic Affairs*
*Hebei Chemical & Pharmaceutical College*
*Shijiazhuang 050000, China*
*E-mail: ml6cc2000@163.com*

*Abstract: Knowledge mapping of medicinal plants enable ordinary people to differentiate between medicinal plants and learn their pharmacological effects, provide assistances and instructions to medical workers during the use of medicinal plants, and support intelligent queries of the properties of traditional medicinal plants. This paper innovatively introduces artificial neural network to the knowledge mapping of medicinal plants, and provides a practical and valuable reference for scientific development and reasonable use of medicinal plants. Firstly, the entity relationships were designed for medical knowledge map, and the definitions, scales, and examples were given for each type of data in the proposed knowledge map of medicinal plants. Next, the authors detailed the ideas of multi-source knowledge fusion, and the acquisition and storage strategies for entity information of medicinal plants. Then, the attention-based bidirectional gated recurrent network was combined with convolutional neural network to detect the genetic relationships between medicinal plants from the angles of semantics and texts. Finally, this paper explains the semantic retrieval algorithm for medicinal plants, and visualizes the knowledge map. The proposed model and semantic retrieval algorithm were proved effective and superior through experiments. It is concluded that: The smaller the batch size, the higher the recognition accuracy of plant entities, and the better the recognition effect. The research findings provide a reference for knowledge mapping in other fields.*

*Keywords: Neural network, Medicinal plants, Knowledge map, Genetic relationship detection, Multisource knowledge fusion.*

## Introduction

China boasts an abundance of medicinal plants. With the development of medicine and agriculture, Chinese scholars have achieved fruitful results on medicinal plants. The relevant research mainly deals with the following issues: component determination, separation, and extraction; pharmacological experiments; development, application, introduction, and cultivation [4, 10, 24]. More than 5,000 medicinal plants, which are wholly or partially medicinal, are recorded in medicine books like Chinese Materia Medica, Medicinal Flora of China, Dictionary of Traditional Chinese Medicine, Pharmacognosy, and Compilation of Chinese Herbal Medicine. The effective parts of the partially medicinal plants include roots, seeds/fruits, flowers/leaves/skins, or vines/stems. Among them, over 200 medicinal plants can be introduced and cultivated [5, 7, 11]. The properties of these diverse medicinal plants need to be described by lots of texts. These plants can be classified by medicinal performance and pharmacological actions. There is a genetic relationship between plants in the same class, i.e., these plants are correlated in terms of chemical composition, pharmacodynamics, and activity [21, 23, 27].

The various medicinal plants form a massive system with complex drug names, confusing pharmacological effects, and fuzzy effect descriptions. For ordinary people with little medical

knowledge, it is an immense challenge to learn medicinal plants without teachers. With the aid of visualization technology, knowledge mapping can describe the correlations between knowledge resources and their carriers, creating graphs that visualize or map the knowledge domain [9, 18, 20]. Knowledge mapping of medicinal plants enable ordinary people to differentiate between medicinal plants and learn their pharmacological effects, guide medical workers to properly use medicinal plants, and support experts and scholars to make intelligent queries of the properties of traditional medicinal plants. The resulting knowledge maps provide a practical and valuable reference for scientific development and reasonable use of medicinal plants [1, 2, 14].

Domestic and foreign scholars have explored the key technologies for the construction and application of knowledge maps in different disciplines. In developed countries, knowledge maps are being applied to more and more fields, yielding good effects [3, 8, 16, 17, 19, 22].

In the field of smart transportation, Kalender et al. [12] plotted a traffic knowledge map by saving time-related dynamic data as attributes and space-related static data as entities. The knowledge map greatly improves the query efficiency of traffic data. In addition, the time and space map convolutional network was adopted to realize independent mining of spatiotemporal correlations, during traffic data prediction.

In the field of finance, Lecue and Wu [15] prepared the knowledge map of financial enterprise information based on knowledge extraction and learning model, and entity similarity algorithm. For knowledge library query, cypher statement was adopted to extract the relevant information through the matching between nodes and relationships on the knowledge map. On this basis, a financial enterprise knowledge service platform was designed, which involves functional modules like display, question and answer (Q&A), library update, and comprehensive authority management. The structure of the platform was described in details.

In the field of medical health, Velampalli and Jonnalagedda [26] extracted medical knowledge from massive multi-source unstructured data through natural language processing (NLP), and designed a multilevel statement division mechanism to convert descriptive texts from files to statements. The mechanism prevents excessively long descriptive texts. Kim [13] extracted knowledge from medical dictionaries and compilations, and constructed a sequence labeling model, which fuses medical knowledge with general knowledge. They also designed a medical entity feature extractor for matching tasks, and a deep matching network. Furthermore, a statement extraction strategy was proposed to extract file-level relationships, based on the filtering of medical entities with the maximum interval.

In the field of military, Franco-Salvador et al. [6] classified the structured and unstructured data crawled from relevant websites, and analyzed their internal correlations, according to the features of these data. On this basis, an ontology model was constructed for the military field. Then, entity classification, entity recognition, and relationship extraction were completed based on k-nearest neighbors (KNN) algorithm, bidirectional long short-term memory-conditional random memory (BiLSTM-CRF), and Pulse Coupled Neural Network (PCNN).

In the field of education, Tsukagoshi et al. [25] summarized several problems with the traditional learning path recommendation systems: low accuracy, poor efficiency, and unreasonable knowledge modeling. To solve these problems, a knowledge map was developed for senior high school learners, and a Q&A system and a knowledge connectivity

state model were designed based on student demands. These tools help the learners to master the knowledge points, which have not been well understood before.

Based on the techniques, experience, and suggestions of experts and scholars about knowledge mapping in various fields, this paper innovatively introduces artificial neural network (ANN) to the knowledge mapping of medicinal plants. Section 2 designs the entity relationships for the knowledge map of medicinal plants, and gives the definitions, scales, and examples for each type of data in the proposed knowledge map. Section 3 partially explains the ideas of multi-source knowledge fusion, and the acquisition and storage strategies for entity information of medicinal plants. Section 4 integrates the attention-based bidirectional gated recurrent network (BGRN) with convolutional neural network (CNN) to detect the genetic relationships between medicinal plants from the angles of semantics and texts. Section 5 explains the semantic retrieval algorithm for medicinal plants, and visualizes the knowledge map. The superiority of our model and semantic retrieval algorithm was demonstrated through experiments.

## Design of entity relationships

The knowledge map of medicinal plants consists of knowledge about the entities and genetic relationships of medicinal plants. Referring to known properties of medicinal plants, this paper designs a knowledge map of 5,118 known medicinal plants, covering 9 types of entities and 7 kinds of genetic relationships. The large map contains sufficient, high-quality, diverse entity knowledge. Tables 1 and 2 give the definitions, scales, and examples for each type of data in the proposed knowledge map.

Table 1. Names, scales, and examples of entities in the knowledge map of medicinal plants

| Name | Meaning | Scale | Example |
|------|---------|-------|---------|
| CN | Chinese name | 3,319 | Platycodon grandiflorum |
| AN | Alternative name | 12,984 | Baggage flower, bell flower, mitral flower |
| CA | Category | 425 | Roots |
| GD | Geographical distribution | 26,423 | China, Korean Peninsula, Japan, and Eastern Siberia |
| GH | Growing habits | 22,164 | Cool weather-grown, cold-resistant, heliophilous |
| FI | Major functions | 18,743 | Diffuse the lung, relieve sore throat, dispel phlegm, expel pus |
| PH | Pharmacological actions | 55,632 | Dispel phlegm and relieve cough; lower blood sugar; diminish inflammation; act on circulation system; suppress gastric liquid secretion, and inhibit ulcer |
| MA | Main prescriptions | 56,275 | Cure lung abscess; cure productive cough and excessive panting; cure pharyngitis and toxin; cure cold repletion chest bind/patients with no heat pattern; cure cold damage, mass at the right hypochondrium, and suffocating sensation in the chest; cure ulcerative gingivitis |
| CA | Clinical applications | 8,943 | Lonicerae and forsythiae powder; mulberry leaf and chrysanthemum beverage; cough-stopping powder |

Table 2. Names, scales, and examples of relationships
in the knowledge map of medicinal plants

| Name | Meaning | Scale | Example |
|---|---|---|---|
| BT | Belongs to | 3,562 | Platycodon grandiflorum, belongs to, roots |
| PC | Plant chemotaxonomy | 12,984 | Platycodon grandiflorum, plant chemotaxonomy label, secondary metabolite |
| MP | Metabolic path | 15,733 | Platycodon grandiflorum, metabolic path, saponins, other metabolic enzymes |
| CC | Chemical composition | 45,362 | Platycodon grandiflorum, chemical composition, volatile oils, fatty acids, saponins, phenolic compounds, etc. |
| PA | Pharmacological activity | 46,537 | Platycodon grandiflorum, pharmacological activity, dispel phlegm, relieve cough, diminish inflammation, inhibit ulcer, suppress nervus centralis |
| GR | Genetic relationship | 18,797 | Genus Mahonia, genetic relationship, Berberis |
| ER | Effect relationship | 25,446 | Mahonia aquifolium, effect relationship, alkaloids |

By defining the entities and genetic relationships of medicinal plants, this paper converts the semi-structured data on medicinal herbs from various sources, such as Chinese Medicinal Plants and Illustrations of Medicinal Plants. After combing through the textual relationships of the data, the descriptive texts of medicinal herbs were compiled into a knowledge map with interconnected genetic relationships. The knowledge map must have sufficient, diverse, and high-quality knowledge on the entities. Therefore, this paper extracts knowledge from the data on medicinal herbs from large industrial websites related to the field of medicine, and the relevant open links, including A Complete Collection of Medicinal Herbs in China (CCMHC), Baidu Baike, zyzyw.com, 39.net, etc.

## Knowledge fusion and information processing

This paper fuses the multi-source knowledge on medicinal plants from two levels: the entity level and the relationship level.

The entity-level fusion mainly targets the plant entity set obtained from the most authoritative source of CCMHC. The entities from other sources, such as Baidu Baike, were added as supplementary information to the set. The plant entities in the set were labeled to facilitate the subsequent expansion of knowledge information.

The relationship-level fusion needs to (1) establish the genetic relationships between the labeled plant entities in the plant entity set, as well as (2) those between the new plant entities from other sources and the labeled plant entities. For the first purpose, the entity relationships in CCMHC data were taken as the primary data, and those in Baidu Baike data as supplementary data; the genetic relationships of plant entities sharing the same name were subjected to knowledge fusion. For the second purpose, the genetic relationships belonging to the new entities from the other sources were firstly added to the knowledge map, such that the new entities are correlated with the entities in the established plant entity set.

The plant entities and their genetic relationships were supplemented and covered by mapping specific knowledge points to the knowledge map nodes, with the highly professional medical

data from CCMHC as the starting point. The new and old entities of the proposed knowledge graph for medicinal plants and their relationships were iterated to compare and fuse new and old knowledge, thereby updating and completing the knowledge map. Fig. 1 shows the structure of the entity information system in our knowledge map for medicinal plants.
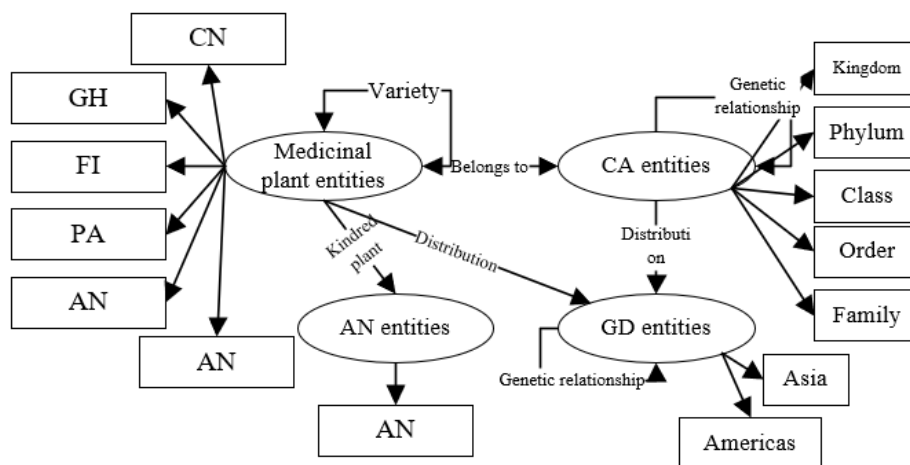


Fig. 1 Structure of the entity information system

Different forms of information about medicinal plant entities should be acquired by different methods along different paths.

The structured information from CCMHC is described and stored in standard formats. The knowledge implied in the relevant texts was transferred by data reading, while the information stored in tables was extracted by professional online crawlers. After analyzing the format of the target webpage, the extraction formats were obtained for such information as CN, AN, PH, and FI. Then, the multimedia data, including texts and images, extracted from the corresponding positions were saved locally.

The medical plant entries on Baidu Baike are semi-structured data. For these entries, the above extraction formats were also adopted to extract the necessary data from the corresponding positions.

The other ordinary webpages offer nonstructured data like descriptive texts. To extract valuable information on the entities and genetic relationship of medicinal plants, the professional terms, e.g., the names and concepts of medicinal plants, should be sorted out before data analysis.

In this way, different approaches were adopted to extract entity information of medicinal plants from the structured, semi-structured, and unstructured data from different sources, and fuse it with the entity information already in the knowledge map, making the map more diverse and complete.

Fig. 2 shows the genetic relationships between medicinal plants of Berberis, including *Mahonia bealei* (Fort.) Carr., *Mahonia bijuga* Hand.-Mazz., *Mahonia bodinieri* Gagnep., *Mahonia bracteolata* Takeda, *Mahonia breviracema* Y. S. Wang et Hsiao, *Mahonia cardiophylla* Ying et Boufford, *Mahonia conferta* Takeda, *Mahonia decipiens* Schneid., *Mahonia duclouxiana* Gagnep., *Mahonia eurybracteata* Fedde, *Mahonia fordii* Schneid., *Mahonia gracilipes* (Oliv.) Fedde, and *Mahonia hancockiana* Takeda.
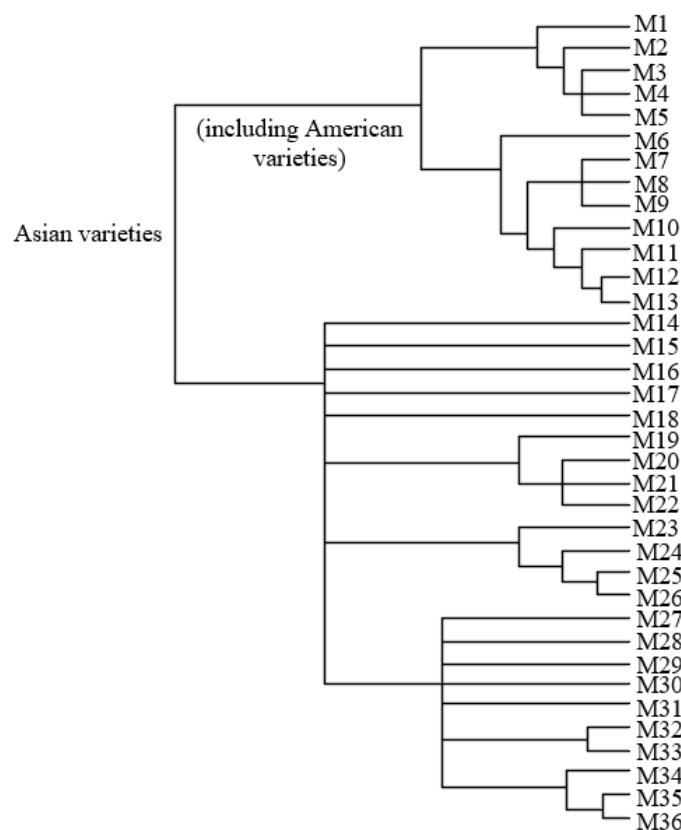
Fig. 2 Genetic relationships between medicinal plants of Berberis

Neo4j Graph Database, which contains nodes, attributes, and relationships, was selected to save the knowledge map of medicinal plants independently. In the database, the nodes represent the entities of medicinal plants, and the edges between nodes stand for the genetic relationships between entities. Each node has a unique label, which specifies its CA, GD, GH, PA, etc. Each edge covers three basic information: start point, destination, and name. Weight and other attributes could be adopted to support the calculation or constrain the knowledge query. The edge information mainly includes the correlations of PC, MP, CC, PH, GR, and ER with the pharmacokinetics and pharmaceutical effects of medicinal plants.

## Genetic relationship detection

The genetic relationship detection of medicinal plants in the knowledge map can be described as follows. After defining the relationship detection problem RD, a set EC of entity candidates is obtained through entity detection, and a set RC of relationship candidates is further derived. Then, the genetic relationship can be predicted based on the matching score $MS\left(RD, r^{i}\right)$ between every genetic relationship in RC and every plant entity:

$$\dot{r}^{i} = \underset{r^{i} \in RC}{argmax}\, MS\left(RD, r^{i}\right). \tag{1}$$

To improve the detection effect of genetic relationships between medicinal plants, the attention-based BGRN was combined with CNN to detect the genetic relationships from the angles of semantics and texts. The detection process mainly includes three steps: mapping relationship candidates into vectors; computing the similarity between vectors by similarity algorithm or deep CNN; judging the semantic similarity between medicinal plants by vector

similarity. Our network model integrates the merits of attention-based BGRN and CNN. Specifically, the sequence modeling function of BGRN was relied on to extract the sematic-level correlations between medicinal plans; the attention mechanism was adopted to track the plant entities and their genetic relationships; the CNN was employed to extract the interaction and matching between words of genetic relationship texts.

Fig. 3 shows the structure of the proposed attention-based BGRN. The model is made up of an input layer, a hidden layer, and an output layer. The input layer vectorizes the input descriptive text of medicinal plants.
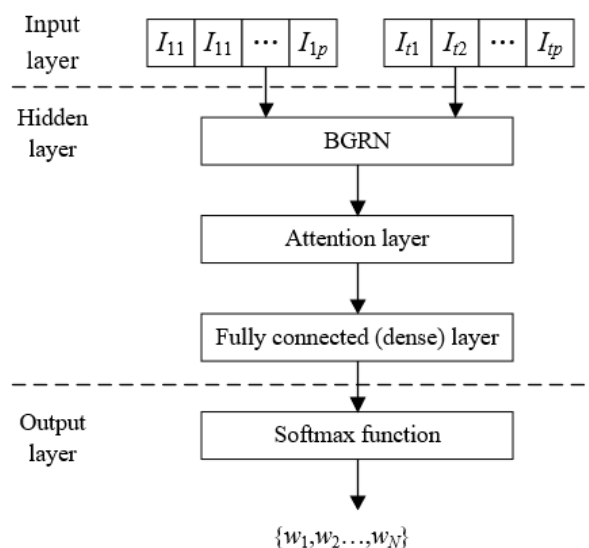


Fig. 3 Structure of attention-based BGRN

The hidden layer encompasses three parts: a BGRN, an attention layer, and a fully connected (dense) layer. Consisting of two gated recurrent networks in opposite directions, the BGRN processes the descriptive text to obtain its word vectors. This network enables the overall consideration of the entire descriptive text during genetic relationship detection. The understanding of the input descriptive text hinges on the weight proportion of different entities in the statement converted from the text. For simplicity, the statement converted from the text is hereinafter referred to as the statement. Therefore, the word vectors of the text outputted by the BGRN were introduced to the attention layer, where each word vector is assigned a probability weight. The weight assignment helps to identify the plant entities and genetic relationships that better convey the meaning of the statement. The fully connected (dense) layer outputs the result of the hidden layer.

The above result needs to be processed by the Softmax function in the output layer. The predicted genetic relationship was mapped to the interval of (0, 1).

Fig. 4 illustrates the structure of our detection model for the genetic relationships of medicinal plants. The RD and granularity of genetic relationships were determined comprehensively from the angles of semantics and texts. The semantic-based text processing was completed by the attention-based BGRN (on the right side of Fig. 4), which outputs features $\rho_1$ and $\rho_2$. The text-based text processing was completed by the CNN coupling a similarity matrix (on the left side of Fig. 4), which outputs features $\rho_3$ and $\rho_4$. The four output features, $\rho_1$, $\rho_2$, $\rho_3$

and $\rho_4$ were fused, and linearly superimposed on the encoder output $\{w_1, w_2, \ldots, w_N\}$ into the desired matching score $MS\left(RD, r^i\right)$.
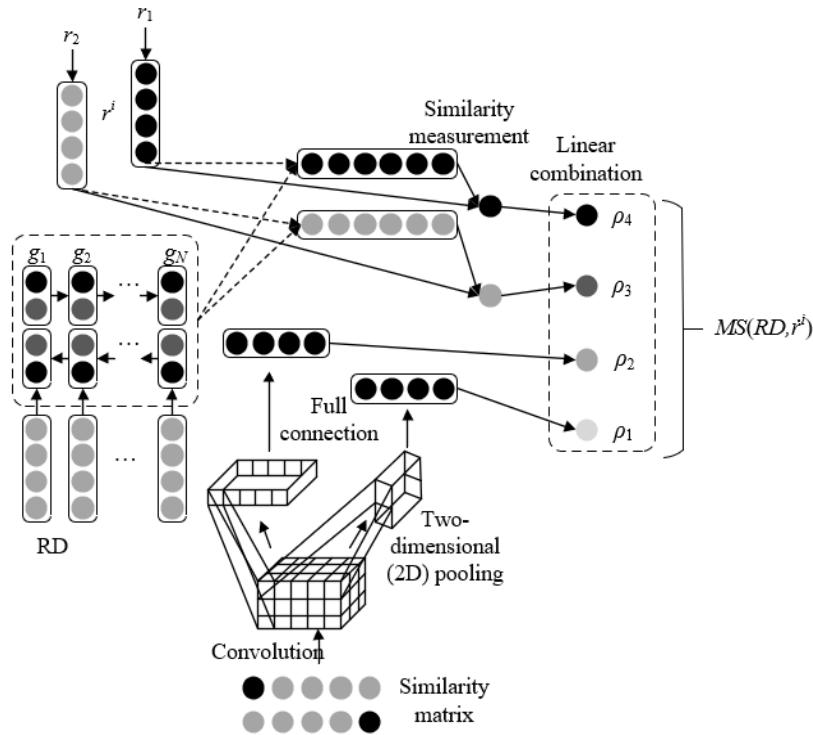


Fig. 4 Structure of genetic relationship detection model

During the semantic-based text processing, the first step is to divide RD into the representation of plant entities and the representation of genetic relationships, and initialize the problem randomly. Next, the embeddable vector of each word in the statement was obtained by an encoder. The connection between the positive and negative hidden representations of an embedded word vector at time k can be denoted as $g_k$, and the set of hidden representations can be expressed as $G_{1:N} = [g_1, g_2, \ldots, g_N]$. The weight vectors, which reflect the degree of importance, determine the actual meaning of the statement. The size of each weight vector depends on the degree of attention in the attention mechanism. Let $\theta_{ik}$ be the attention weight of the $k$-th word in the statement; $N$ be the number of words in the statement. Let $X$ and $Y$ be the dimensions of $r$ and $g$, respectively; $a$ be a hyperparameter. Then, the dimensions of the parameters $\delta$ and $\lambda$ to be updated can be expressed as $a \times (X + Y)$ and $1 \times a$, respectively. Therefore, the RD problem can be described as:

$$RD_i = \sum_{k=1}^{N} \theta_{ik} g_k = \sum_{k=1}^{N} \frac{e^{\omega_{ik}}}{\sum_{k=1}^{N} e^{\omega_{ik}}} g_k = \sum_{k=1}^{N} \frac{e^{\delta^{\mathrm{T}} tan\left(\lambda^{\mathrm{T}} [g_k ; r_k] + \varepsilon\right)}}{\sum_{k=1}^{N} e^{\delta^{\mathrm{T}} tan\left(\lambda^{\mathrm{T}} [g_k ; r_k] + \varepsilon\right)}} g_k \, , \tag{2}$$

where $i = 1$ represents RD of plant entities; $i = 2$ represents the characterization of genetic relationships. Let $\circ$ be the product operation between vectors. Then, the similarity can be calculated by:

$$S_i = RD_i \circ g_i . \tag{3}$$

The correlation on the text level should also be contemplated to fully consider the descriptive text, i.e., the RD problem should be treated as a basic text matching problem. Let $e_i$ be the $i$-th word representing the plant entity in the statement; $c_i$ be the $j$-th word representing genetic relationship in the statement. Then, a similarity matrix SM can be established to reflect the basic interactions between plant entities and genetic relationships:

$$SM_{ij} = e_i \bullet c_j, \tag{4}$$

where $\bullet$ is the operation of similarity measurement using the cosine function. The matrix design differentiates our algorithm from strict text matching algorithm, such as to extract the similarity between words in different forms.

The CNN extracts the convolution features from the match patterns of words or characters on different levels and in different sequences. The feature map $I^l$ can be obtained by scanning the SM with the $l$-th kernel $K^l$:

$$I_{i,j}^l = \operatorname{Re} LU \left( \sum_{x=0}^{q^l-1} \sum_{y=0}^{q^l-1} K_{x,y}^l \cdot SM_{i+x, j+y} + \varepsilon^l \right), \tag{5}$$

where $q^l$ is the size of kernel $K^l$ and Re$LU$ is the activation function of rectified linear unit.

Then, two pooling layers of the size $1 \times b_1$ and $1 \times b_2$ were adopted for max pooling of the feature map $I^l$:

$$\begin{cases} o_i^{(1,l)} = \max_{0 \le h \le b_1} I_{i,h}^l \\ o_j^{(2,l)} = \max_{0 \le h < b_2} I_{h,j}^l \end{cases} \tag{6}$$

Through the convolution and pooling by the CNN, the best matching scores between each word and each relationship and between different words were obtained from the world level for the RD problem. Suppose there are $M$ kernel functions. Then, the cascaded output of the $M$ kernel functions in max pooling can be denoted as $[o^{(i, 0)}, o^{(i, M)}]$. The results of two max pooling operations were processed by the two perceptron's in the fully connected layer. The final output features can be expressed as:

$$\begin{cases} \rho_3 = \sigma_2 \cdot \operatorname{Re} LU \left( \sigma_1 \left[ o^{(1,0)}, o^{(1, M)} \right] + \varepsilon_1 \right) + \varepsilon_2 \\ \rho_4 = \sigma_2 \cdot \operatorname{Re} LU \left( \sigma_1 \left[ o^{(2,0)}, o^{(2, M)} \right] + \varepsilon_1 \right) + \varepsilon_2 \end{cases}, \tag{7}$$

where $\sigma_i$ is the weight coefficient of the $i$-th perceptron. The activation function is still Re$LU$.

Through the semantic- and text-level matching of the descriptive text by the attention-based BGRN and the CNN, four eigenvalues were obtained: $\rho_1$, $\rho_2$, $\rho_3$, and $\rho_4$. The four eigenvalues were then linearly combined to derive the final matching score $MS(RD, r^i)$:

$$MS(RD, r^i) = Sigmoid \left( \sigma^{\mathrm{T}} [\rho_1, \rho_2, \rho_3, \rho_4] + \varepsilon \right). \tag{8}$$

The above model was trained, and loss ranking was performed to maximize the marginal income of the positive relationships $r^+$ and other relationships $r^-$ in the set $R$ of relationship candidates:

$$L\left(RD, \hat{r}^{*}, r^{*}\right) = \sum_{\left(RD, \hat{r}^{*}\right) \in D} \max\left(0, \eta + MS\left(RD, \hat{r}^{*}\right) - MS\left(RD, r^{*}\right)\right), \qquad (9)$$

where $\eta$ is a constant in the interval of [0, 1].

## Semantic retrieval and knowledge map visualization

Each entity in the knowledge map has a wealth of knowledge about the corresponding medicinal plant. The retrieval statement of map users also contains the information about the plant entities and genetic relationships. To determine the plant corresponding to each entity in the map, it is necessary to compute the similarity between the retrieval statement and the plant entities in the map.

Fig. 5 explains the flow of semantic extraction from the knowledge map of medicinal plants.
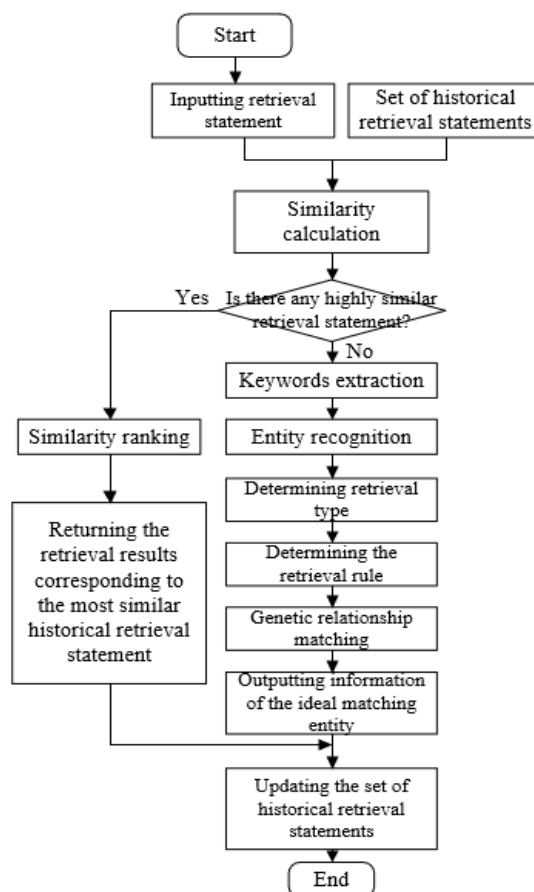


Fig. 5 Flow of semantic extraction

Firstly, the similarity between the input retrieval statement and the set of historical retrieval statements is calculated, and the retrieval result corresponding to the most similar historical retrieval statement is returned. If the similarity calculation fails to achieve a desirable result, the retrieval statement will subject to entity identification and keywords extraction. The entity identification aims to match the name of medicinal plant in the retrieval statement with the

plant entities in the knowledge map. To extract the keywords, the first step is to determine the retrieval type of the retrieval statement: entity retrieval, relationship retrieval, or combined retrieval. Then, the retrieval rule is determined based on the retrieval type. After that, the ideal matching entity will be outputted, according to the matching results between the entity information in the retrieval statement and the entities/genetic relationships in the knowledge map. Finally, the retrieval statement with undesired matching result is added to the set of historical retrieval statements, and the information of the set is updated.

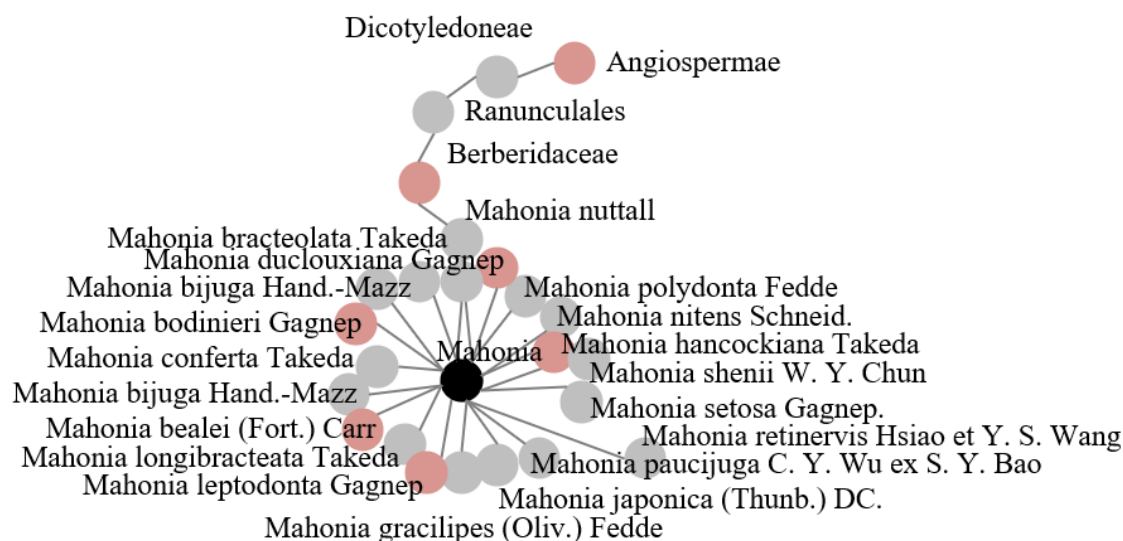The proposed knowledge map is vitalized by Neo4j technology (Fig. 6).



Fig. 6 Visualized knowledge map

## Experiments and results analysis

From the perspective of application effect, our model matches the descriptive text on medicinal plants in RD from semantic and text levels, and thereby cuts down the number of candidate relationships.

Fig. 7 presents the recall-precision curves on candidate set, which is plotted on MATLAB, and their equilibrium points. It can be seen that the two-level matching (curve 1) achieved a larger area under the curve (AUC) and a higher equilibrium point than semantic-based matching (curve 2) and text-based matching (curve 3). Therefore, the proposed two-level matching can generate a better candidate set, which contains lots of medicinal plants closely related to the retrieval entity.

The number of iterations was set to 50, 100, 150, 200, 250, 300, 350, 400, 450, and 500 in turn to test the extraction effect of the BGRN on the semantic-level genetic relationships of medicinal plants.

Table 3 records the precisions at different number of iterations. Between the 400th and 500th iterations, the precision maintained at around 87-88%, and tended to converge. There was virtually zero oscillation of precision, which remained at about 88%. To verify the performance of our BGRN in relationship extraction, our model was compared with a bidirectional LSTM, with words and characters as the basic units. The comparison results (Table 4) show that the two models were more precise when characters were taken as the

basic units; whether words or characters served as the basic units, our model outperformed the bidirectional LSTM. Therefore, our model is both effective and feasible.
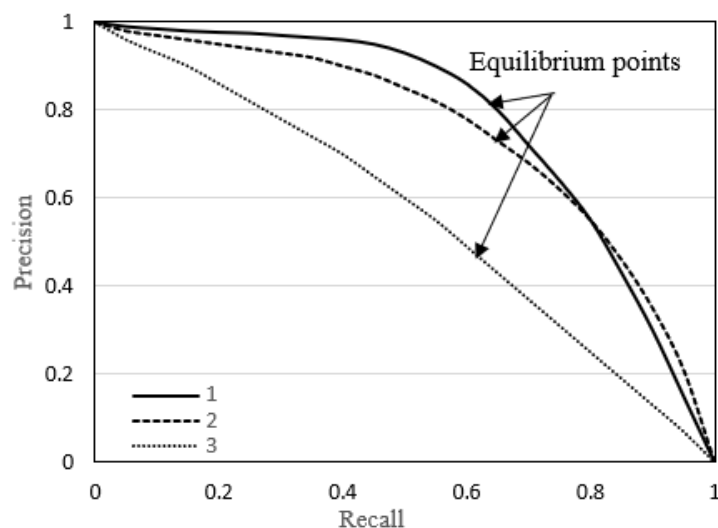


Fig. 7 Recall-precision curves on candidate set

Table 3. Precisions at different number of iterations

| Number of iterations | Precision | Number of iterations | Precision |
|---|---|---|---|
| 50 | 34% | 300 | 73% |
| 100 | 43% | 350 | 80% |
| 150 | 47% | 400 | 87% |
| 200 | 52% | 450 | 87% |
| 250 | 67% | 500 | 88% |

Through systematic performance testing on our knowledge map of medicinal plants, it was learned that the retrieval effect of our model depends heavily on the batch size. Therefore, Fig. 8 compares the precision convergence of our model under changing batch size. It can be inferred that the retrieval precision of our model increased gradually with the growing number of iterations, and eventually tended to be stable. The smaller the batch size, the higher the recognition accuracy of plant entities, and the better the recognition effect.

To further verify the knowledge retrieval effect of the proposed knowledge map, several models were designed, namely, LSTM, character-level enhanced encoder (CLEE), world-level recurrent neural network (WRNN), attention-based CNN (ACNN), and hierarchical residual learning (HRS)-based LSTM (HLSTM), and compared through a retrieval test based on simple questions. The test results of these models and our model are compared in Table 5. The correctness of a retrieval result depends on whether the predicted genetic relationship between medicinal plants is consistent with the actual relationship. It can be seen that our model, i.e., attention-based BGRN + CNN, achieved better retrieval accuracy than all the other model on different manual databases.

Table 4. Test results on genetic relationship extraction

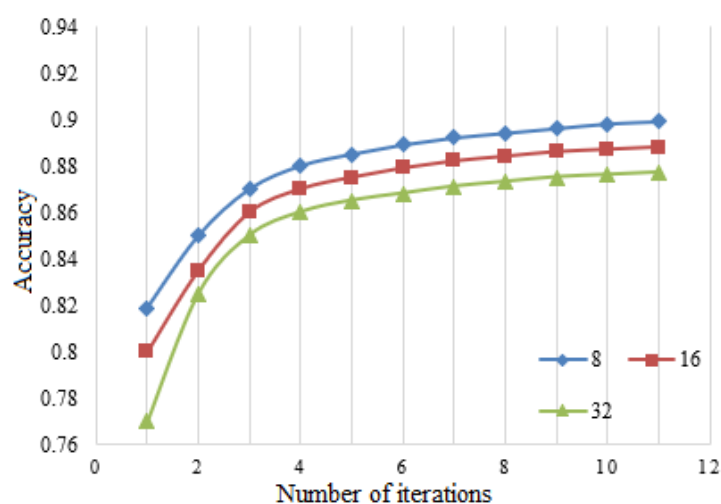| Unit | Model | Serial number | Precision | Model | Serial number | Precision |
|---|---|---|---|---|---|---|
| Words | Bidirectional LSTM | 1 | 75% | Our model | 11 | 87% |
| | | 2 | 66% | | 12 | 72% |
| | | 3 | 68% | | 13 | 81% |
| | | 4 | 58% | | 14 | 72% |
| | | 5 | 74% | | 15 | 78% |
| | | 6 | 62% | | 16 | 89% |
| | | 7 | 69% | | 17 | 81% |
| | | 8 | 59% | | 18 | 73% |
| | | 9 | 74% | | 19 | 81% |
| | | 10 | 77% | | 20 | 75% |
| | | Mean | 68.2% | | Mean | 78.9% |
| Characters | Bidirectional LSTM | 21 | 72% | Our model | 31 | 75% |
| | | 22 | 71% | | 32 | 76% |
| | | 23 | 70% | | 33 | 79% |
| | | 24 | 68% | | 34 | 82% |
| | | 25 | 69% | | 35 | 88% |
| | | 26 | 77% | | 36 | 81% |
| | | 27 | 56% | | 37 | 72% |
| | | 28 | 72% | | 38 | 79% |
| | | 29 | 74% | | 39 | 85% |
| | | 30 | 71% | | 40 | 88% |
| | | Mean | 70% | | Mean | 80.5% |



Fig. 8 Precision convergence curves at different batch sizes

The processed genetic relationships between medicinal plants were saved in the local Neo4j Graph Database. Map users need to perform retrieval in Cypher language. The database will return each user the information about the retrieval plant entity and the entities that are genetically related to that entity. Fig. 9 is an overview of the proposed knowledge map, which reflects the knowledge connections of root-type dicotyledoneae medicinal plants.

Table 5. Retrieval accuracies of different models

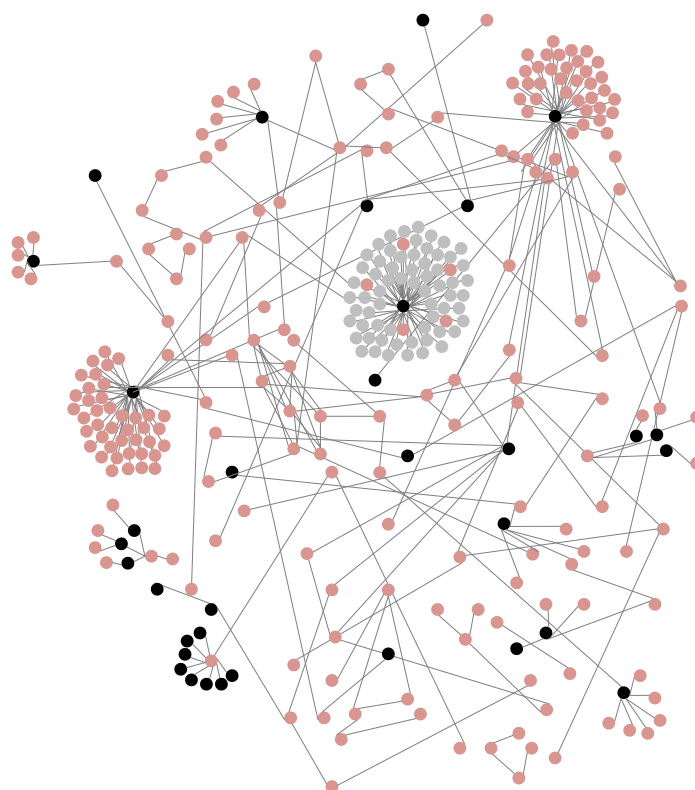| Model | Database 1 | Database 2 |
|---|---|---|
| LSTM | 65.4% | 66.8% |
| CLEE | 71.3% | 72.8% |
| WRNN | — | 70.5% |
| ACNN | 78.1% | — |
| HLSTM | 76.2% | — |
| Our model | 81.9% | 83.8% |



Fig. 9 Overview of our knowledge map

## Conclusions

This paper mainly constructs an ANN-based knowledge map for medicinal plants. Firstly, the entities and relationships of the knowledge map were designed, and the definitions, scales, and examples were given for each type of data in the map. Next, the authors explained the ideas of multi-source knowledge fusion, as well as the way to acquire and store the information of plant entities. With the aid of attention-based BGRN and CNN, the genetic relationships of medicinal plants were detected from the angles of semantics and text. Finally, a semantic retrieval algorithm was developed for medicinal plants, followed by an overview of the knowledge map. From the perspective of application effect, contrastive experiments were designed. The experimental results show that: (i) whether words or characters served as the basic units, our model outperformed the contrastive model; (ii) the smaller the batch size, the higher the recognition accuracy of plant entities, and the better the recognition effect; (iii) our model achieved better retrieval accuracy than other models.

This paper innovatively develops a knowledge mapping system for medicinal plants, and details the construction of the genetic relationship detection module. The entities,

relationships, and properties of medicinal plants were defined systematically based on the existing data. The research findings provide a prescription guide for traditional Chinese medicine practitioners, promote the intelligent disease diagnosis and medication, and facilitate the transform of scientific knowledge map into professional medical knowledge map.

## References

1. Aasman J. (2017). Transmuting Information to Knowledge with an Enterprise Knowledge Graph, IT Professional, 19(6), 44-51.
2. Abdelaziz I., A. Fokoue, O. Hassanzadeh, P. Zhang, M. Sadoghi (2017). Large-scale Structural and Textual Similarity-based Mining of Knowledge Graph to Predict Drug-drug Interactions, J Web Semant, 44, 104-117.
3. Ait-Mlouk A., L. Jiang (2020). KBot: A Knowledge Graph Based chatBot for Natural Language Understanding over Linked Data, IEEE Access, 8, 149220-149230.
4. Bakhshi M., M. Nematbakhsh, M. Mohsenzadeh, A. M. Rahmani (2020). Data-driven Construction of SPARQL Queries by Approximate Question Graph Alignment in Question Answering over Knowledge Graphs, Expert Syst Appl, 146, 113205.
5. Franco-Salvador M., P. Gupta, P. Rosso, R. E. Banchs (2016). Cross-language Plagiarism Detection over Continuous-space- and Knowledge Graph-based Representations of Language, Knowledge-based Systems, 111, 87-99.
6. Franco-Salvador M., P. Rosso, M. Montes-y-Gómez (2016). A Systematic Study of Knowledge Graph Analysis for Cross-language Plagiarism Detection, Information Processing & Management, 52(4), 550-570.
7. Futia G., F. Morando, A. Melandri, L. Canova, F. Ruggiero (2015). ContrattiPubblici.org, A Semantic Knowledge Graph on Public Procurement Information, In AI Approaches to the Complexity of Legal Systems, 380-393.
8. Gardasevic S. (2020). User-driven Efforts in Creating Knowledge Graph Information System, Digital Library Perspectives, 36(2), 97-111.
9. Hasan S. S., D. Rivera, X. C. Wu, E. B. Durbin, J. B. Christian, G. Tourassi (2020). Knowledge Graph-enabled Cancer Data Analytics, IEEE J Biomed Health, 24(7), 1952-1967.
10. Ienco D., R. G. Pensa (2019). Enhancing Graph-based Semisupervised Learning via Knowledge-aware Data Embedding, IEEE T Neur Net Lear, 31(11), 5014-5020.
11. Johansson J., M. Contero, P. Company, F. Elgh (2018). Supporting Connectivism in Knowledge Based Engineering with Graph Theory, Filtering Techniques and Model Quality Assurance, Adv Eng Inform, 38, 252-263.
12. Kalender M., M. T. Eren, Z. Wu, O. Cirakman, S. Kutluk, G. Gultekin, E. E. Korkmaz (2018). Videolization: Knowledge Graph Based Automated Video Generation from Web Content, Multimed Tools Appl, 77(1), 567-595.
13. Kim H. (2017). Towards a Sales Assistant Using a Product Knowledge Graph, J Web Semant, 46, 14-19.
14. Kumar A. J., C. Schmidt, J. Köhler (2017). A Knowledge Graph Based Speech Interface for Question Answering Systems, Speech Communication, 92, 1-12.
15. Lecue F., J. Wu (2017). Explaining and Predicting Abnormal Expenses at Large Scale Using Knowledge Graph Based Reasoning, J Web Semant, 44, 89-103.
16. Le-Phuoc D., H. N. M. Quoc, H. N. Quoc, T. T. Nhat, M. Hauswirth (2016). The Graph of Things: A Step towards the Live Knowledge Graph of Connected Things, J Web Semant, 37, 25-35.
17. Martinez-Rodriguez J. L., I. Lopez-Arevalo, A. B. Rios-Alvarado (2018). OpenIE-based Approach for Knowledge Graph Construction from Text, Expert Syst Appl, 113, 339-355.

18. Minkov E., K. Kahanov, T. Kuflik (2017). Graph-based Recommendation Integrating Rating History and Domain Knowledge: Application to On-site Guidance of Museum Visitors, J Assoc Inf Sci Tech, 68(8), 1911-1924.

19. Pira E., V. Rafe, A. Nikanjam (2018). Searching for Violation of Safety and Liveness Properties Using Knowledge Discovery in Complex Systems Specified through Graph Transformations, Inform Software Tech, 97, 110-134.

20. Rahman M. M., A. Takasu (2020). Leveraging Entity-type Properties in the Relational Context for Knowledge Graph Embedding, IEICE Transactions on Information and Systems, 103(5), 958-968.

21. Rezig S., Z. Achour, N. Rezg (2016). Control Synthesis Based on Theory of Regions with Minimal Reachability Graph Knowledge, IFAC-PapersOnLine, 49(12), 1383-1388.

22. Seo S., B. Oh, K. H. Lee (2020). Reliable Knowledge Graph Path Representation Learning, IEEE Access, 8, 32816-32825.

23. Solihin W., C. M. Eastman (2016). A Knowledge Representation Approach in BIM Rule Requirement Analysis Using the Conceptual Graph, J Inf Technol Constr, 21(21), 370-401.

24. Song M., W. Zhao, E. HaiHong (2020). KGAnet: A Knowledge Graph Attention Network for Enhancing Natural Language Inference, Neural Comput Appl, 32(18), 14963-14973.

25. Tsukagoshi Y., T. Kawamura, Y. Sei, Y. Tahara, A. Ohsuga (2020). Proposal of Knowledge Graph and Completion Method for Solving Social Issues: An Attempt to Improve the Bicycle Parking Environment in the University, IEEJ Transactions on Electronics, Information and Systems, 140(8), 905-915.

26. Velampalli S., M. V. Jonnalagedda (2017). Graph Based Knowledge Discovery Using MapReduce and SUBDUE Algorithm, Data Knowl Eng, 111, 103-113.

27. Verborgh R., M. Vander Sande, O. Hartig, J. Van Herwegen, L. De Vocht, B. De Meester, P. Colpaert (2016). Triple Pattern Fragments: A Low-cost Knowledge Graph Interface for the Web, J Web Semant, 37, 184-206.

**Lei Miao**

E-mail: ml6cc2000@163.com

Lei Miao is a lecturer at Office of Academic Affairs of Hebei Chemical & Pharmaceutical College, Shijiazhuang, China. He is mainly engaged in machine learning, image processing, and the application of artificial neural network in image recognition.