Journal of Combinatorial Mathematics
and Combinatorial Computing
www.combinatorialpress.com/jcmcc

Combinatorial Press

Research on the association between acupuncture points and diseases based on the construction of knowledge graph and applied research

Yue Liu[1,✉], Yunzhi Zhang[2]

[1] *Hainan Vocational University of Science and Technology, Haikou, Hainan, 571101, China*
[2] *The 928th Hospital of People's Liberation Army Joint Logistic Support Force, Haikou, Hainan, 571101, China*

ABSTRACT

Acupuncture has been recognized by more and more experts as a treatment method to relieve various pains in human body, but the association between specific acupuncture treatments and diseases is still unclear, which affects the long-term development of acupuncture treatment. In this paper, we abstract the knowledge of acupuncture points as ontologies in the knowledge graph, and propose a method to improve the RoBERTa-WWM-BiGRU-CRF model to optimize the knowledge extraction of the knowledge graph by combining the SoftLexicon technique and the adversarial training method. Based on the knowledge graph of acupuncture points, the collaborative filtering model is introduced, and the original similarity matrix construction method is replaced by the co-occurrence matrix construction method based on the association characteristics of acupuncture points and diseases, which improves the operational efficiency of the association search and realizes the design of the association search technology of acupuncture points and diseases. The average consultation time in the acupuncture outpatient departments of the experimental and control groups applying this paper's technology for acupuncture visits was faster than that of the full outpatient clinic by 0.32 min, showing a significant difference ($P<0.05$). Patients in the experimental group who received acupuncture treatment assisted by the technology of this paper were higher than those in the control group in the dimensions of acupuncture treatment experience, such as physiological reflections, treatment emotions, and treatment effects and treatment feeling dimensions, which were 2.22, 3.57, 2.2, and 1.33, respectively.

Keywords: knowledge graph, acupuncture points, SoftLexicon technology, collaborative filtering modeling

✉ Corresponding author.
E-mail address: hkd2022138011@126.com (Y. Liu).
Received 05 August 2024; Revised 25 October 2024; Accepted 15 February 2025; Published Online 16 April 2025.
DOI: 10.61091/jcmcc127b-187

1. Introduction

Acupuncture and moxibustion is an important part of traditional Chinese medicine (TCM), a representative of external treatment method of TCM, and a crystallization of the wisdom of the Chinese nation. It utilizes needling or heat moxibustion of specific parts of the human body surface in order to achieve symptom relief or disease treatment [1]. As a complementary and alternative therapy, acupuncture has been confirmed by a large number of clinical observations and experimental studies for its comprehensive regulatory effects on various systems or organs of the body, and is now widely used for a wide range of health care, disease prevention, and symptom relief [2-4].

Despite the abundance of literature related to acupuncture, there are fewer findings of regularity among them, and most of them are centered on the disease of concern, and the establishment of a data network related to the acupoints is information [5-6]. Acupuncture points are the body's internal organs and meridians gas and blood infusion in and out of the body surface of the special parts of the body, both gas and blood convergence, transfer and access to specific premises, but also with the internal organs and meridians of the gas connected and with the activities, changes in the feeling points, conduction points and treatment points, with the feeling of the stimulus and reflecting the evidence of the disease of the two major functions [7]. Modern research has confirmed that in pathological conditions, acupuncture points are out of the activation state, and the function is present in the pathological state [8-9]. For example, when the internal organs of the human body are locally damaged or when a disease occurs, the corresponding skin acupoint area reflects the changes in the strength and size of the area of the sensitive points of the acupoints, which is closely related to the severity of the corresponding disease damage [10-12]. The use of existing acupoint databases has revealed that current databases generally record the grouping of acupoints related to diseases, and the general rule of synergistic or antagonistic effects of acupoint pairings is less explored [13-14]. In acupuncture and moxibustion, acupoint prescription is the key link between preventive health care and treatment of diseases, so the establishment of a knowledge graph-based database of acupuncture point and disease associations is an important scientific issue for the modernization and development of acupuncture and moxibustion [15-17].

In the face of the growing demand for knowledge of acupuncture points and disease association, this paper firstly designs a knowledge map of acupuncture points relying on the relevant data of Chinese medicine acupuncture points, formulates the construction process of the knowledge map of acupuncture points, and completes the data collection and preprocessing work. In the knowledge extraction of acupuncture point knowledge graph, oriented to the characteristics of acupuncture point data, we propose a TCM named entity recognition model based on BIO annotation method and combining SoftLexicon and adversarial training to improve the RoBERTa-WWM-BiGRU-CRF model to realize the ternary representation of knowledge. Adversarial training FGM is introduced to generate adversarial samples by applying small perturbations to the word vectors to enhance the generalization ability and robustness of the model and to finalize the knowledge fusion and storage of the knowledge graph. In this paper, the research focuses on the collaborative filtering model based on the knowledge graph of acupuncture points, which realizes the association relationship calculation and association search process on the data of acupuncture points and disease-related information. The item-based collaborative filtering model was then modified to calculate the similarity between acupuncture points and disease content. Aiming at the problem of high computational complexity of the similarity values in the model similarity matrix, the method of constructing the similarity matrix by replacing the formula of cosine similarity used in the original model with the construction method of co-occurrence matrix is proposed, which reduces the computational complexity in constructing the similarity matrix and improves the operational efficiency of the whole algorithm. The performance of this paper is examined by carrying out the test of acupuncture point and disease association searching technology from a number of indexes such as checking accuracy rate, recall rate, F1 value, MAE, and so on. Finally,

the technology of this paper is applied to the real acupuncture diagnosis and treatment work, and the practical significance and role of this paper's technology is explored through the application of practical analysis.

2. Knowledge Mapping of Acupuncture Points and Disease Associations

The knowledge system in the field of acupuncture points is complex, and the problem of knowledge fragmentation is serious, knowledge graph can be used to solve these problems by well-structured preservation and extension of knowledge can be used to solve these problems [18]. In this chapter, we first determine the construction process of acupuncture point knowledge graph, propose a method combining SoftLexicon technology and adversarial training to improve the RoBERTa-WWM-BiGRU-CRF model to extract knowledge from text more efficiently, and finally, complete the knowledge fusion and knowledge storage of the acupuncture point knowledge graph to establish the acupuncture point knowledge graph for the later acupuncture Finally, the knowledge fusion and knowledge storage are completed to establish the knowledge graph of acupuncture points, which will provide data support for the research of acupuncture points and disease association and the association search technology in the following paper.

2.1. Knowledge graph construction process

The construction process of acupuncture point knowledge map includes ontology construction in the schema layer, and knowledge extraction, knowledge fusion and knowledge storage in the data layer. In the schema layer, the ontology construction is based on the existing acupuncture point knowledge atlas, relevant Chinese medicine standards, relevant literature and other information to refine the acupuncture point knowledge, and the schema layer is designed to define the concepts of entities, attributes and inter-entity relationships. In the data layer, deep learning model is used to extract knowledge automatically in the knowledge extraction phase, then different sources of knowledge are integrated and unified in the knowledge fusion phase, and finally the knowledge is stored in the Neo4j graph database to populate the data into the knowledge graph.

2.2. Data collection and pre-processing

The basis of constructing knowledge graph lies in the huge data set, in which data collection is the important support of the whole data layer construction process, and the quality of the data set determines the specialization and reliability of the knowledge graph in a specific domain.

Although there is a lot of information about medical treatment on the Internet, there is less information about acupuncture points. In order to ensure the completeness and richness of the dataset, the acupuncture points dataset consists of the following three parts. One part is from the online medical encyclopedia website “Medical.com”, which is a very professional Chinese medical encyclopedia in China, edited by national encyclopedia enthusiasts and reviewed by domestic professionals to provide users with medical knowledge services. Another part from the “acupuncture and moxibustion therapeutics”, this part of the data is the main content of the disease and its acupuncture point guidance, specifically contains the content of the disease, dialectical points, matching points of treatment, etc.; the last part from the “Chinese medicine disease classification and code”, mainly contains a specific description of the concepts of Chinese medicine diseases, symptoms and treatments.

2.3. Knowledge extraction based on RoBERTa-WWM-BiGRU-CRF modeling

The dataset related to acupuncture points is relatively small, and the use of traditional manual and rules for knowledge graph construction is not only costly in terms of labor, but also unfavorable for knowledge expansion of the knowledge graph at a later stage. The following research will focus on the

knowledge extraction work that is crucial in the knowledge graph construction task of acupuncture points.

2.3.1. Knowledge extraction methods. In order to realize automatic knowledge extraction and reduce the manual cost in constructing the knowledge graph, BIO annotation is performed on the acquired dataset, and the combination of SoftLexicon and adversarial training to improve the RoBERTaWWM-BiGRU-CRF named entity recognition model is proposed, as well as the training and prediction on the homemade dataset [19].

Combining the above works, an automatic knowledge extraction method for acupuncture point knowledge graph based on combining SoftLexicon and adversarial training to improve RoBERTaWWM-BiGRU-CRF named entity recognition model is designed.

2.3.2. RoBERTa-WWM-BiGRU-CRF entity recognition model. The SRFBC model contains four core components: an embedding layer, an adversarial training layer, a BiLSTM layer, and a CRF layer.

1) Embedding Layer. In order to solve the problem of entity boundary recognition in the acupuncture point domain, in the embedding layer, each word in the text sequence is first converted to the corresponding word vector by the RoBERTa-WWM pre-trained language model to capture the word features. At the same time, the SoftLexicon method is used to introduce the lexical information associated with each word into the word vector, and this word fusion mechanism enables the character-level vectors to encompass a wider range of semantic attributes.The RoBERTa-WWM model is based on BERT, and it has been significantly improved for the two pre-training tasks, namely, MLM (Masked Language Modeling) and NSP (Next Sentence Prediction).

The input text of the model is defined as $S=\{c_1,c_2,c_3,c_4,...,c_n\}$. The character vector x_i^c for each character in the text can be obtained by RoBERTa-WWM embedding as shown in Eqs. (1) and (2):

$$x_i^c = e^c(c_i) \tag{1}$$

$$x_i^c = \left\lfloor e^c(c_i); e^b(c_i, c_{(i+1)}) \right\rfloor \tag{2}$$

where: e^c represents the index table of character vectors; e^b represents the index table of character vectors of bigram.

With the SoftLexicon method, each character is identified in the input sequence s by cross-referencing it with the lexicon to identify all matching words. Then, these matches are categorized into four lexical sets {B, M, E, S}: $B(c_i)$ includes all words starting with character c_i, $M(c_i)$ contains words with character c_i in the middle position, $E(c_i)$ covers words ending with c_i, and $S(c_i)$ is a lexical set consisting of a separate c_i. In this way, each character is given information about its position in the word. These four word sets are shown in equation (3):

$$\begin{cases} B(c_i)=\{w_{i,k}, \forall w_{i,k}\in L, i<k,, n\} \\ M(c_i)=\{w_{j,k}, \forall w_{j,k}\in L, 1,, j<i<k,, n\} \\ E(c_i)=\{w_{j,i}, \forall w_{j,i}\in L, 1<j,, i\} \\ S(c_i)=\{c_i, \exists c_i\in L\} \end{cases} \tag{3}$$

where: L represents the dictionary; w is the word from the input text sequence starting at c_i back to c_k back and appearing in the dictionary L. If there is no word in the dictionary that matches the structure of {B, M, E, S}, the empty set representation is used. After recognizing that each character belongs to one of the four categories of the {B, M, E, S} lexicon, these lexicons are converted into fixed-length vectors. The conversion process utilizes word frequencies as weights to perform a weighted

average, which is used to generate a representative vector for the word set S. The computation is detailed in equation (4).

$$\begin{cases} v^s(S) = \dfrac{4}{Z}\sum_{w\in S} z(w)e^w(w) \\ Z = \sum_{w\in B\cup M\cup E\cup S} z(w) \end{cases} \tag{4}$$

where: $z(w)$ is the frequency of word w in the dictionary; Z is the accumulation of word frequencies in the word set; and $v_s(S)$ denotes the processed set vector. After converting the word sets into vectors, the vectors of the four word sets are merged into the vector representation of each character by the vector splicing method to form the ultimate representation of each character as shown in Eq. (5):

$$\begin{cases} e^s(B,M,E,S) = \left\lfloor v^s(B), v^s(M), v^s(E), v^s(S) \right\rfloor \\ x^c \leftarrow \left[x^c; e^s(B,M,E,S) \right] \end{cases} \tag{5}$$

After completing the encoding of text feature vectors, in order to address the generalization ability and robustness of the model on the acupuncture point dataset, the model introduces an adversarial training FGM. Due to the above introduction of SoftLexicon method to realize word fusion mechanism and embedding of text via RoBERTaWWM.

2) Adversarial Training. FGM is an efficient method for generating adversarial samples.FGM first calculates the gradient of the model's loss function with respect to the embedding layer for the input data under the current parameters. This gradient indicates the direction in which the loss function increases the fastest. The FGM then adds a small perturbation to the embedding layer along the direction of this gradient to generate adversarial samples. Considering making the perturbation most effective as well as time efficient, the choice here is to add the perturbation in the direction where the loss increases the most. This is because adding a perturbation in this direction maximizes the impact on the model's prediction results and thus improves the robustness of the model more effectively. The calculation is shown in equation (6):

$$D[\max_{\Delta x\in\Omega} L(x+\Delta x, y;\theta)] \tag{6}$$

where: D represents the training set; x represents the inputs; y represents the labels; θ is the model parameters; $L(x,y;\theta)$ is the loss of a single sample; Δx is the adversarial perturbation; Ω is the perturbation space; and max below $\Delta x\in\Omega$ denotes the finding of the loss function that makes the loss function L maximized Δx. In order to maximize the adversarial perturbation and to prevent it from Δx being too large as shown in equations (7) and (8):

$$r_{abv} = e\frac{g}{P\,g\,P_2} \tag{7}$$

$$g = \nabla_x L(\theta, x, y) \tag{8}$$

In Eq. (8), r_{abv} represents the maximum perturbation value i.e. Δx; ε represents the intensity factor; g represents the gradient of the loss function with respect to the output of the BERT embedding layer x.

3) BiGRU layer. BiGRU is an algorithm that combines bi-directional RNN and GRU, which utilizes forward and reverse GRU units to process the input sequences in depth and splices the outputs from these two directions to fully capture the contextual information in the sequences [20].

In the model the input sequence is $X=\{x_1,x_2,x_3,...,x_T\}$. The forward computation steps of BiGRU model are as follows:

$$\overrightarrow{h_t}=GRU(\overrightarrow{h_{t-1}},x_t) \tag{9}$$

$$\overleftarrow{h_t}=GRU(\overleftarrow{h_{t+1}},x_t) \tag{10}$$

where, Eq. (9) and Eq. (10) are the hidden states computed from left to right by the forward GRU and from right to left by the reverse GRU, respectively, for the time step (t). GRU stands for GRU unit, and x_t stands for the t nd element of the input sequence. The forward and reverse hidden states are spliced to obtain the final hidden state h_t as shown in equation (11):

$$h_t=\left[\overrightarrow{h_t};\overleftarrow{h_t}\right] \tag{11}$$

Finally the spliced hidden state h_t is passed to a fully connected layer and the output y_t is obtained by softmax function as shown in equation (12):

$$y_t=softmax(Wh_t+b) \tag{12}$$

where: W & b are the weights and bias of the fully connected layer.

4) CRF layer. Conditional Random Field (CRF), a discriminative probabilistic undirected graph model, is commonly used for tasks such as sequence labeling. In CRF, the relationship between input sequences and output sequences is usually represented using a feature function, where the weights of the feature function are obtained through learning. The optimal feature function weights are obtained by maximizing the log-likelihood function and training the CRF model using methods such as gradient descent. In named entity recognition tasks, there are usually some constraints between labels, such as BIO annotation rules.

In the named entity recognition task, a linear chain conditional random field is usually used, and given an input sequence of $X=\{x_1,x_2,x_3,\ldots,x_n\}$ and an output sequence of $Y=\{y_1,y_2,y_3,\ldots,y_n\}$, the conditional probabilities when X takes the value of x and Y takes the value of y can be used in Eqs. (13) and (14):

$$P(y\mid x)=\frac{1}{Z(x)}\exp\left(\sum_{i,k}\lambda_k t_k\left(y_{i-1},y_i,x,i\right)+\sum_{i,l}\mu_l s_l\left(y_i,x,i\right)\right) \tag{13}$$

$$Z(x)=\sum_{y}\exp\left(\sum_{i,k}\lambda_k t_k\left(y_{i-1},y_i,x,i\right)+\sum_{i,l}u_l s_l\left(y_i,x,i\right)\right) \tag{14}$$

where: t_k,s_l is the characteristic function; λ_k,u_l is the corresponding weights; $Z(x)$ is the normalization factor.

2.3.3. Experimentation and analysis. In this section, the same data source as constructing the knowledge graph is chosen as the main experimental data resource, and the performance of this paper's knowledge extraction model based on RoBERTa-WWM-BiGRU-CRF is verified through experiments. In the knowledge extraction task, recall and F1 value are adopted as the evaluation indexes, which can comprehensively reflect the performance of the model as well as facilitate comparison and communication with other models.

In the performance test experiments, the Word2Vec-based model, TextTCNN-based model, FastTexT-based model, BERT-based model and this paper's model are compared under the same conditions, and the results are shown in Table 1. As can be seen from the table, the knowledge extraction performance of this paper's model is significantly better than other traditional commonly

used knowledge extraction models, and there is also a certain progress than the BERT-based model, the precision rate, the recall rate and the F1 value reach 84%, 84.1%, 83.3%, respectively.

Table 1. Experimental result

Model	Precision	Recall	F1
Word2Vec	0.687	0.685	0.694
TextTCNN	0.747	0.737	0.74
FastTexT	0.749	0.746	0.753
BERT	0.795	0.78	0.78
Model of this article	0.84	0.841	0.833

2.4. Knowledge integration

Due to the wide range of knowledge sources and the diversity of Chinese textual expressions, knowledge extraction may lead to information redundancy, and as a key link in knowledge reprocessing, the main role of knowledge fusion is to improve the consistency and credibility of information. Entity alignment is the most critical work of knowledge fusion, which aims to unify nodes with the same semantics. It is difficult to formulate more complete knowledge fusion rules for this kind of situation, this paper adopts the cosine similarity method based on word frequency statistics to calculate the similarity between entities, and screens the entity pairs whose similarity is higher than the threshold by setting a suitable threshold, which is calculated as follows:

$$S_{(m_1,m_2)} = \frac{\sum_{i=1}^{n}(A_i B_i)}{\sqrt{\sum_{i=1}^{n}(A_i)^2}\sqrt{\sum_{i=1}^{n}(B_i)^2}} \tag{15}$$

where m_1, m_2 represents two entities; A_i and B_i are the i th dimension elements in the word frequency vectors obtained after the disambiguation of m_1 and m_2, respectively; and n is the number of disambiguations. The cosine similarity method based on word frequency statistics realizes the mapping of word vectors from semantic space to vector space and evaluates the entity similarity based on the cosine value between the vectors, i.e., the larger the cosine value, the more similar the two entities are. After several experimental tests, the best effect is achieved when the threshold is set to 0.6.

2.5. Knowledge storage

Emergency scenario text data is formed into structured emergency scenario triples after data processing, knowledge extraction and knowledge fusion steps. In this paper, the data is converted into CSV format and the Py2neo library is used to batch import the data into the Neo4i graph database to realize the storage of the emergency scenario triples.

3. Acupuncture point and disease association search technique

Among all kinds of biomedical data, the most closely related data to human beings are disease-related data. Whether we can better utilize the knowledge map of acupuncture points and more conveniently grasp the association between acupuncture points and diseases is of great significance to improve the efficiency of acupuncture treatment and enhance the research level of the association between acupuncture points and diseases. In this chapter, based on the knowledge map of acupuncture points, we will study the principle of collaborative filtering and modify the collaborative filtering model to

realize the calculation of the association relationship between acupuncture points and diseases and the association search process on the disease-related information data.

3.1. Collaborative Filtering Model

Collaborative filtering algorithms have become a hotspot in the research of recommender systems due to their concise design concept and superior computational performance [21]. In most cases, when users shop and don't know what products to buy, they seek advice from relatives, friends or classmates, and the collaborative filtering algorithm applies this idea. It does this by analyzing the user's past behavior, establishing a link between the user and the product, and recommending the product to the user based on the opinions of other users. Generally, those customers who have similar preferences in the past will have similar preferences in the future. Collaborative filtering algorithms are basically categorized into domain-based and model-based. Among them, domain-based collaborative filtering can be categorized into user-based collaborative filtering algorithms and item-based collaborative filtering algorithms based on different objectives. The implementation of collaborative filtering algorithms includes three main steps: collecting user information, finding similar items, and calculating and generating recommendations.

1) Collecting user information and building user models. Collecting user preference information is divided into two forms: explicit and implicit, explicit is to collect the user's direct evaluation of the project, and implicit is by collecting the user's time to browse the page, the number of clicks, and the viewed information. There are also some data in the collected information that are noisy, which may be caused by misoperation or disconnection, etc., and there are also some missing data caused by users' unwillingness to evaluate, which will have an important impact on the quality of the recommendation. Therefore, it is necessary to reduce noise and fill the data. Under normal circumstances, the number of user browsing is greater than the number of user purchase records, in order to take the value in a reasonable range of values, so that the collected information is more accurate, the use of normalization, which ensures that different user behavior in a reasonable range of values.

2) Finding Nearest Neighbors. By calculating the similarity between the target user and other users, find similar neighbors and generate the target user's nearest neighbor set. Commonly used similarity measures in collaborative filtering algorithms are Euclidean distance, cosine similarity, modified cosine similarity and Pearson correlation coefficient.

(1) Euclidean distance. Euclidean distance calculates the distance between any two points in the space vector, which is only related to the coordinates of the two points. The smaller the Euclidean distance, the greater the similarity. The formula is shown below:

$$d(x,y)=\sqrt{\sum_{i=1}^{k}(x_i-y_i)^2} \tag{16}$$

$$sim(x,y)=\frac{1}{1+d(x,y)} \tag{17}$$

where: x_i denotes the coordinates of point x, y_i denotes the coordinates of point y, $sim(x,y)$ denotes the similarity between point x and point y, and $d(x,y)$ denotes the distance between the two points.

(2) Cosine similarity. The similarity is calculated by calculating the cosine angle between the vectors. The lower the cosine value, the closer the two vectors are. The formula for cosine similarity is as follows:

$$sim(x,y)=\frac{|N(x)\cap N(y)|}{\sqrt{|N(x)||N(y)|}} \tag{18}$$

where: $N(x)\cap N(y)$ denotes the set of items that have been fed back by both user x and user y, and $|N(x)||N(y)|$ denotes the total number of items that have been fed back by user x and user y.

(3) Modified cosine similarity. The modified cosine similarity takes into account the differences in ratings between different users, and reduces the computational error by introducing the average of user ratings on top of the cosine similarity to regulate the rating deviation vector. The modified cosine similarity calculation formula is as follows:

$$sim_{a\cos}(u,v)=\frac{\sum_{i\in I_{uv}}\left(r_{u,i}-\overline{r_v}\right)}{\sqrt{\sum_{i\in I_u}\left(r_{u,i}-\overline{r_u}\right)^2}\sqrt{\sum_{i\in I_v}\left(r_{v,i}-\overline{r_v}\right)^2}} \tag{19}$$

where $I_{u,v}$ is the common set of rated items for user u and user v, r_{ui} is used to denote user u's rating of item i, $r_{v,i}$ is used to denote user v's rating of item i, $\overline{r_u}$ is the average rating of user u, and $\overline{r_v}$ is the average rating of user v.

(4) Pearson correlation coefficient. Pearson's correlation coefficient measures the similarity between users based on their common rating items and is calculated as follows:

$$sim_{pearson}(i,j)=\frac{\sum_{c\in I_y}\left(R_{i,c}-\overline{R_i}\right)\left(R_{j,c}-\overline{R_j}\right)}{\sqrt{\sum_{c\in I_y}\left(R_{i,c}-\overline{R_i}\right)^2}\sqrt{\sum_{c\in I_y}\left(R_{j,c}-\overline{R_j}\right)^2}} \tag{20}$$

where R_i, denotes the rating of item c by user i, $\overline{R_i}$ and $\overline{R_j}$ denote the average rating of the item by user i and user j, respectively, and I_{ij} is the set of items jointly rated by user i and user j.

3.2. Collaborative Filtering Optimization Model

Collaborative filtering models include user-based collaborative filtering algorithms and item-based collaborative filtering algorithms.

User-based collaborative filtering algorithms are user-based recommendation algorithms that discover users' preferences for items based on their historical information and recommend products to users with the same preferences.

Compared with user-based algorithms, project-based collaborative filtering algorithms are more widely used because the growth of projects is different from the growth of users, which is relatively slower and makes the system more stable in comparison. Item-based collaborative filtering algorithm is to find items with similar ratings to the target item, find the nearest neighbor and then predict the target user's rating of the target item, and finally select the top N with the highest ratings for recommendation.

In this paper, we believe that some ideas and calculation methods in the recommendation model based on collaborative filtering can also be used to calculate the association relationship of disease-related information in this paper, and compared with the user-based collaborative filtering model, the item-based collaborative filtering model is more suitable to be applied to the search process of the association between acupuncture points and diseases in this paper. The details are as follows.

1) Calculate the similarity between acupuncture points and diseases. Let there are n recommended contents in the whole system, when calculating the similarity between these n contents, a similarity matrix is used to store these similarities, and let the similarity matrix be W, then Wij represents the similarity between content j and content i. There are various methods to calculate the similarity between data, such as the commonly used Euclidean distance similarity, Jaccard coefficient and so on. In the item-based collaborative filtering model used in this thesis, according to the data relationship abstracted in the above figure and the characteristics of this model, a more suitable similarity calculation method for this thesis is the cosine similarity, whose calculation formula is as follows:

$$W_{ij} = \frac{|N(i) \cap N(j)|}{|N(i)|} \tag{21}$$

where N(i) represents the set of users who have a preference for content i, N(j) represents the set of users who have a preference for content j, the above formula can be interpreted as the proportion of users who like content i who also like content j. There is an obvious problem with the above formula, i.e., if the item j is more popular, and a lot of users like it, then W_{ij} will be very large, and is too close to 1. Thus the formula causes any content to be very similar to the popular content j. To avoid this disadvantage, in practice the following formula is usually used, which penalizes the weight of content j:

$$W_{ij} = \frac{|N(i) \cap N(j)|}{|N(i) \cap N(j)|} \tag{22}$$

2) Generate a recommendation list for the user based on the similarity between acupuncture points and disease content. After obtaining the content similarity matrix, the collaborative filtering process can be carried out. For a target recommendation user u, the model will calculate the degree of interest of user u in content j by the following formula:

$$P_{uj}u = \sum_{i \in N(u)} W_{ij} * R_{ui}i \tag{23}$$

Among them, Puj represents the final rating of user u on content j, N(u) represents the content that user u has generated historical behavior, and Rui represents the strength of user u's previous historical behavior on content i (e.g., the degree of liking, rating, etc.). Analysis of formula (4-3) can be summarized in the project-based collaborative filtering model in the Rank mechanism of the core idea: the user for a content j of the overall rating comes from j and the user's favorite content i of the similarity of the user's favorite content i of the degree of the combined calculation. The significance of synergy lies in the fact that the Rank mechanism needs to be heavily based on the user's historical evaluation of multiple contents when rating a certain content.

3) Generating Recommendations. After obtaining the nearest neighbors of the target user, predict the unrated items of the target user based on the scores of the nearest neighbors, and recommend the top N items with the highest predicted scores to the target user. The prediction formula is as follows:

$$PO(n^2)_{u,c} = \overline{R_u} + \frac{\sum_{v \in N(u)} sim(u,v)(R_{v,c} - \overline{R_v})}{\sum_{v \in N(u)} |sim(u,v)|} \tag{24}$$

where $R_{u,c}$ is the predicted rating value of item c by target user u, $N(u)$ is the set of neighboring users of target user u, and $sim(u,v)$ is the similarity between users.

3.3. Associative Search Algorithm Design Based on Collaborative Filtering Modeling

The recommendation process of the item-based collaborative filtering model was introduced in the previous section, but to use the whole model for the computation of the association relationship in the search process of this paper also needs to map the concepts of certain acupuncture point knowledge graphs in the model and transform some of the processes, so that the transformed model can be more adapted to the data of the disease-related information used in this paper. Among them, the mapping of the model concepts are mainly the following two kinds.

1) User role mapping to disease ontology. In the original collaborative filtering model, the calculation process is mainly to calculate the similarity between the content and the user. In this paper, the part of the data that needs to be calculated is the value of the correlation between the disease-related information and the acupuncture points, and these correlations are mainly composed of the disease ontology and other disease-related information besides the disease. Therefore, it is natural to map the user role to the disease ontology.

2) Mapping the user's historical behavioral information on the content to the association between the disease ontology and other disease information. In the item-based collaborative filtering model, the user's history of content behavior is mainly used as a kind of statistical data to support the Rank process in the second step, after mapping the user's history of content behavior to the correlation between the disease ontology and other disease information, the statistical significance of the model can still be maintained, and there is no computational change in the subsequent Rank process.

3.4. Optimization of association search algorithm based on collaborative filtering model

In this paper, after completing the initial implementation of the search algorithm based on the project-based collaborative filtering model, a series of tests were conducted, and it was found that the time consumption of the similarity calculation and the construction of the similarity matrix between the disease-related data and the data related to the knowledge atlas of acupuncture points was relatively high in the whole search process. There are two computationally intensive processes in the whole algorithm, and the construction of similarity matrix is one of them. Therefore, improving the construction speed of similarity matrix is very helpful to improve the running speed of the whole algorithm.

In this paper, the construction method of co-occurrence matrix is applied in the similarity matrix calculation process of this algorithm, replacing the calculation method of similarity matrix value in Equation (22) [22].

Assuming that the size of the nodes in the attribute graph of the knowledge graph of acupuncture points constructed based on the keywords is n, and the size of the edges representing the correlation relationship between the data in n nodes is m, it is found that if the similarity between acupuncture points and diseases is calculated directly one by one, the time complexity is high because of repeated traversal of the node information and querying of the relationship between the nodes. To complete the construction of a n th-order square matrix, we need to traverse the matrix n^2 times, and when calculating the ith content $N(i)$ in the square matrix, because the data are put into the attribute graph in Graphframes in advance, the computation time of $N(i)$ can be regarded as $O(l)$, and so the overall complexity of completing the construction of the similarity matrix is $W_{ij} = \frac{|N(i) \cap N(j)|}{|N(i) \cap N(j)|}$. When constructing the similarity matrix, the time complexity of the whole construction process is $O(m)$. Since the bipartite graph is sparse in most cases in the item-based collaborative filtering model, the value of m is roughly proportional to that of n, and is much lower than that of n^2.

4. Acupuncture Points and Disease Associated Search Technique Testing

In this chapter, the performance test of the proposed acupuncture point and disease association search technique based on acupuncture point knowledge graph will be carried out. The effectiveness of this paper's acupuncture point and disease association search technique will be verified by comparing the MAE, accuracy and recall of the traditional association rule mining algorithm and Slope One algorithm. The experimental data were obtained from the online medical encyclopedia website "Medicine.com", and the experimental dataset of acupuncture point and disease association was established.

4.1. Comparative analysis of detection rates

The simulation results of this paper's algorithm with traditional association rule mining algorithm and Slope One algorithm in terms of checking accuracy index are shown in Fig. 1. From the simulation results in the figure, it can be seen that the recommendation accuracy of this paper's algorithm is significantly improved with the increase in the number of neighbors input for training, and it is always higher than the traditional association rule recommendation algorithm and Slope One algorithm. When the number of neighbors is 6, the checking accuracy of this paper's algorithm and traditional association rule mining algorithm and Slope One algorithm all reach the highest level, but the checking accuracy of this paper's algorithm reaches 66.5%, which is 14.15% and 4.58% higher than that of traditional association rule mining algorithm and Slope One algorithm.

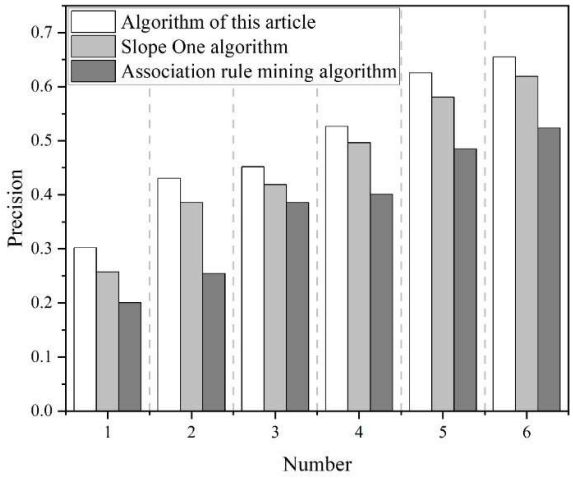


Fig. 1. Precision

4.2. Comparative analysis of recall rates

The simulation results of recall comparison of different algorithms are specifically shown in Fig. 2. The simulation results show that the recall of this paper's algorithm and the traditional association rule mining algorithm and Slope One algorithm do not change significantly as the number of neighbors increases. In general, people always hope that both accuracy and recall increase with the increase of the number of neighbors, however, this is only an ideal state. Although this paper's algorithm does not show a significant increase in recall with the number of neighbors, the recall is always higher than the traditional association rule mining algorithm and Slope One algorithm as a comparison. When the number of neighbors is 30, the recall of this paper's algorithm reaches the lowest 47.08%, but it is still higher than the traditional association rule mining algorithm and Slope One algorithm by 14.93% and 4.61% respectively.

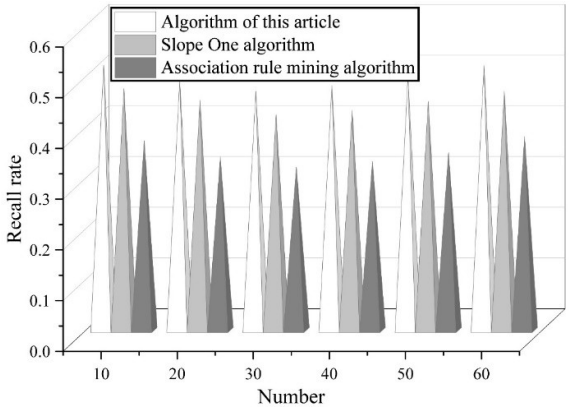


Fig. 2. Recall rate

4.3. Comparative analysis of F1 values

The comparison of F1 values of each algorithm is specifically shown in Fig. 3. A, B and C in the figure correspond to this paper's algorithm, Slope One algorithm and traditional association rule mining algorithm respectively. From the figure, we can see that the diversity index of this paper's algorithm is significantly higher than that of traditional association rule mining algorithm and Slope One algorithm. When the number of neighbors is 4, the F1 value of this algorithm is very similar to that of Slope One algorithm, with a difference of only 1.87%. However, when the number of neighbors grows to 5, the difference between the F1 value of this paper's algorithm and Slope One algorithm expands, and after that the growth rate of the F1 value of this paper's algorithm increases further, and the difference with the traditional association rule mining algorithm, Slope One algorithm expands further. When the number of neighbors is 6, the F1 value of this paper's algorithm reaches the highest 62.11%. The simulation results show that the F1 value of this paper's algorithm, i.e., the diversity index, is significantly higher than that of the traditional association rule mining algorithm and Slope One algorithm, and it can give more accurate results of the association between acupuncture points and diseases.

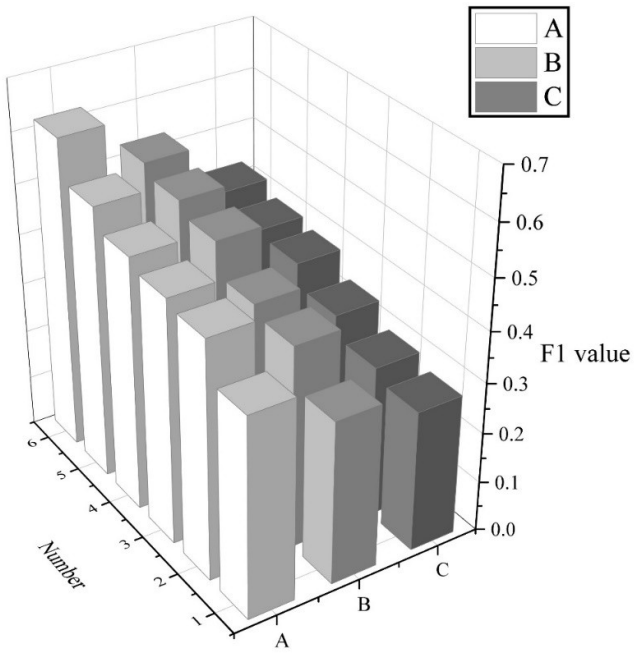


Fig. 3. F1 value

4.4. Comparative analysis of MAE

In this section, the Mean Absolute Error (MAE) metric will be used to measure the accuracy of each algorithm in the association of acupuncture points with diseases.The lower the MAE value, the more accurate the association of acupuncture points with diseases. The comparative simulation results of MAE of this paper's algorithms with traditional association rule mining algorithms and Slope One algorithm are specifically shown in Fig. 4. The MAE values of all algorithms decrease with the increase of the number of neighbors. Compared with traditional association rule mining algorithm and Slope One algorithm, the MAE value of this paper's algorithm is always the lowest. When the number of neighbors is 60, this paper's algorithm reaches the lowest 0.64, compared with the traditional association rule mining algorithm, Slope One algorithm is lower 0.08, 0.04.

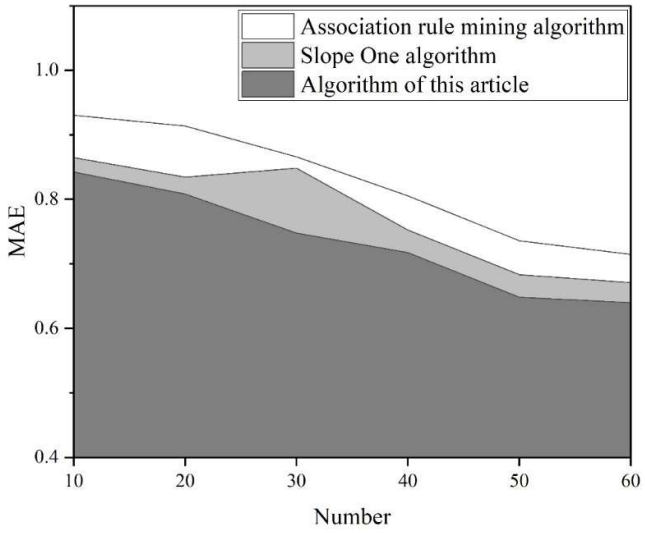
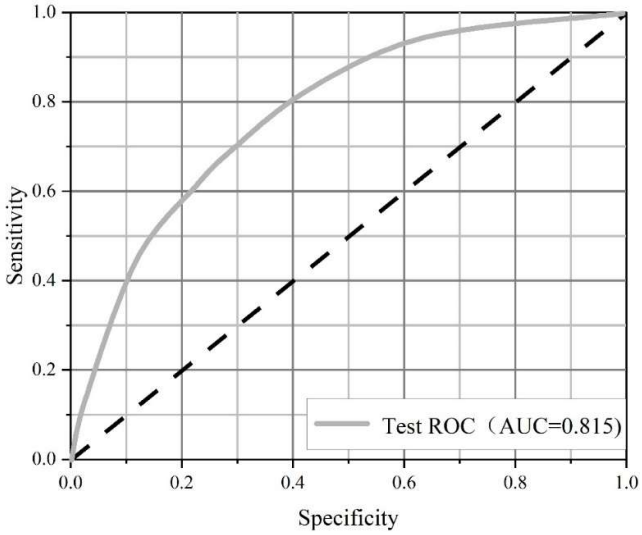


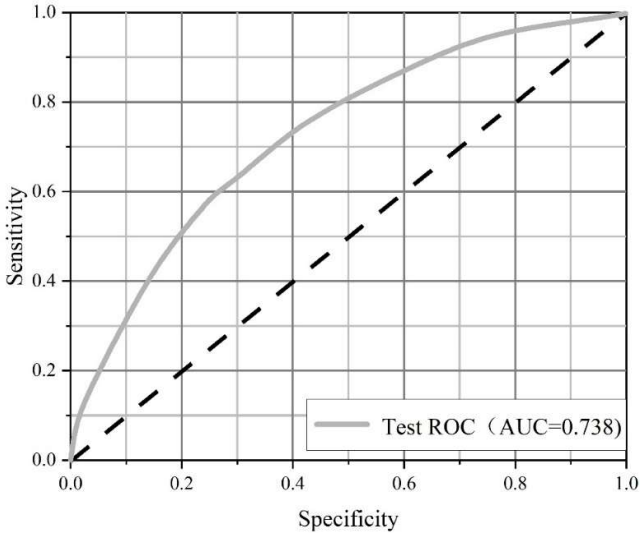
Fig. 4. MAE

4.5. Comparative analysis of ROC curves

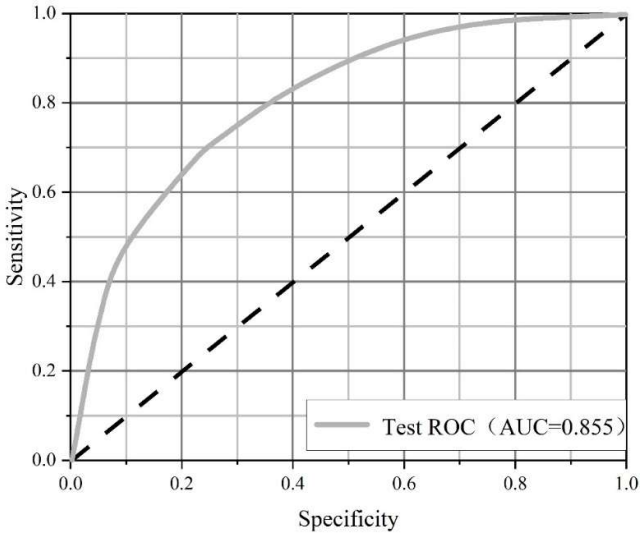
The association results of acupuncture points-diseases of this paper's algorithm with traditional association rule mining algorithm and Slope One algorithm were imported into SPSS20.0 software, and the ROC curves were used to judge the performance of the algorithms. The ROC curves of each algorithm are specifically shown in Fig. 5, and Figs. (a), (b), and (c) are the Slope One algorithm, traditional association rule mining algorithm and this paper's algorithm, respectively. As can be seen from the figure, the areas under the ROC curves of this paper's algorithm and traditional association rule mining algorithm and Slope One algorithm are 0.855, 0.738 and 0.815 respectively. The algorithm in this paper has the largest area under the ROC curve, the highest accuracy of acupuncture point-disease association, and outperforms the other compared algorithms.



(a)Slope One algorithm



(b)Association rule mining algorithm



(c)Algorithm of this article

Fig. 5. ROC curve

5. Acupuncture Points and Disease Associated Search Technique Application Practice

The proposed acupuncture point and disease association search technique based on the acupuncture point knowledge graph facilitates acupuncture practitioners to query the association relationship between acupuncture points and diseases when studying and working, and reduces work pressure. In this chapter, the acupuncture clinic of a traditional Chinese medicine hospital will be used as the research object to carry out the application of the acupuncture point-disease association search technique for 15 days, and to explore whether it has real utility in acupuncture consultation and treatment work.

5.1. Analysis of Acupuncture Visiting Practices

During the practice of acupuncture consultation, the acupuncture clinic will use the acupuncture point and disease association search technology of this paper to assist the doctor's acupuncture consultation work. At the end of the practice, the statistics and integration of acupuncture outpatient department doctor's consultation efficiency, and will be compared with the whole hospital outpatient consultation efficiency, the comparison results are shown in Table 2. From the data in the table, it can be seen that the average consultation time of doctors in the acupuncture clinic is 6.565min, comparing with the whole hospital outpatient clinic is the average consultation time faster by 0.32min, showing a significant difference ($P<0.05$). Obviously, the use of this paper acupuncture point and disease association search technology to assist acupuncture consultation work, can effectively improve the efficiency of doctors' acupuncture consultation.

Table 2. Diagnosis efficiency

Department	Duration of diagnosis(min)		
	General appointment	Sepcialist appointment	Total
Acupuncture clinic	6.08	7.05	6.565
Full outpatient clinic	6.45	7.32	6.885
T	32.826	26.827	53.108
P	<0.05	<0.05	<0.05

5.2. Analysis of Acupuncture Therapy Practices

In this study, patients with frozen shoulder will be used as subjects, and 60 subjects will be recruited according to the inclusion and exclusion criteria, and will be divided into the experimental group and the control group in a randomized and blinded manner. The experimental group will be treated with the proposed acupuncture point and disease association search technique based on the acupuncture point knowledge graph. The control group will maintain the original acupuncture treatment method. At the end of the acupuncture treatment, the acupuncture experience data of 60 patients were collected.

The acupuncture treatment experience of patients in the experimental group and the control group after the acupuncture treatment practice is shown in Figure 6. In the figure, D1 to D4 represent the four dimensions of physiological reflection, therapeutic emotion, therapeutic effect and therapeutic experience in turn, while E and C stand for the experimental and control groups respectively. It can be clearly seen that the mean values of physiological reflection, therapeutic emotion, therapeutic effect and therapeutic experience of patients in the experimental group were higher than those of the control group, with the mean values higher than those of the control group by 2.22, 3.57, 2.2 and 1.33, respectively. The results show that applying the acupuncture point-disease association search

technique in this paper to acupuncture treatment can better clarify the association relationship between the patient's disease and the acupuncture points designed in the corresponding adopted acupuncture treatment, and provide a better acupuncture treatment experience for the patients.

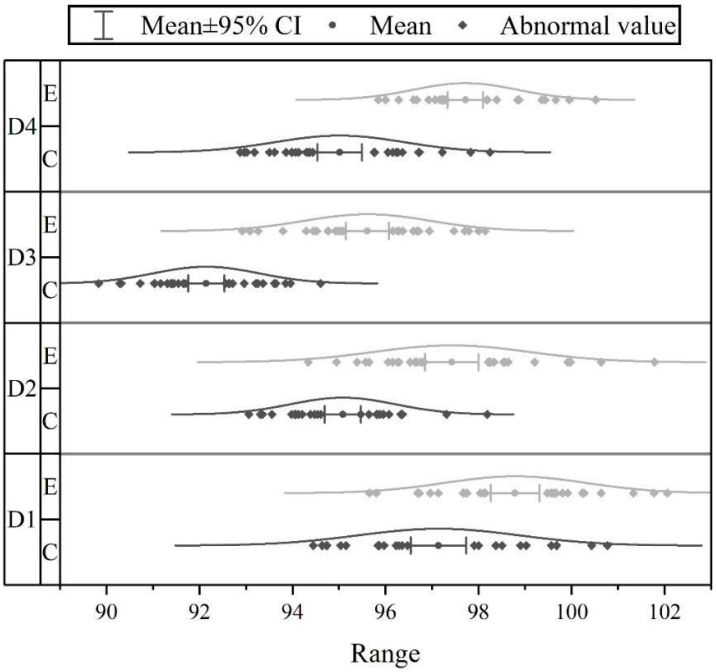


Fig. 6. Acupuncture treatment experience

6. Conclusion

In this paper, we constructed a knowledge graph of acupuncture point and disease associations, and modified the collaborative filtering model on its basis to propose a search technique for acupuncture point and disease associations, which can provide assistance for acupuncture education and acupuncture treatment.

The performance of the proposed acupuncture point and disease association search technique is tested by selecting traditional association rule mining algorithm and Slope One algorithm as a comparison in terms of several indexes, such as detection rate, recall rate, F1, and MAE. In terms of search accuracy, recall, and F1 indexes, comparing with other algorithms, this paper's algorithm always maintains the highest level as the number of neighbors grows and changes. When the number of neighbors is 6, the highest check accuracy rate and F1 value can reach 66.5% and 62.11% respectively. In terms of recall rate, this paper's algorithm shows obvious fluctuation with traditional association rule mining algorithm and Slope One algorithm. When the number of neighbors is 6, the recall rate of this paper's algorithm reaches the lowest 47.08%, but it is still higher than Slope One algorithm, traditional association rule mining algorithm 4.61%, 14.93%. Comparing with Slope One algorithm and traditional association rule mining algorithm, the MAE value of this paper's algorithm is always the lowest, and the MAE value reaches the lowest 0.64 when the number of neighbors is 60, which is lower than that of Slope One algorithm and traditional association rule mining algorithm, 0.08 and 0.04. Judging the performance of algorithms by using the ROC curve, the area under the ROC curve of this paper's algorithm is 0.855, which is higher than Slope One algorithm and traditional association rule mining algorithm. Overall, this paper's algorithm performs well in many indexes such as search accuracy, recall, F1, MAE, etc. This paper's acupuncture point and disease association search technology has good performance.

This paper's acupuncture point and disease association search technology is applied to the acupuncture course acupuncture consultation and treatment work to explore its practical role. In the acupuncture consultation practice, the average consultation time of the acupuncture clinic that applied the technology of this paper to assist the consultation work was 6.565 min, which was faster than that of the whole hospital outpatient clinic by 0.32 min, showing a significant difference ($P<0.05$). In the acupuncture treatment practice, the mean values of physiological reflection, therapeutic emotion, therapeutic effect and therapeutic feeling dimensions in the acupuncture treatment experience of the experimental group were higher than those of the control group, and the mean values were higher than those of the control group by 2.22, 3.57, 2.2 and 1.33, respectively.Overall, the techniques in this paper can improve the efficiency of acupuncture visits in the acupuncture outpatient department of Chinese medicine hospitals in the acupuncture treatment practice, and to a certain degree, enhance and optimize patients' acupuncture treatment experience. This paper's acupuncture point and disease association search technique is of great significance in both acupuncture consultation and treatment.

Funding

This work was supported by Production and education integration characteristic professional group of Hainan Vocational University of Science and Technology (NO. HKZYQ2024-05).

References

[1] Gong, Y., Li, N., Lv, Z., Zhang, K., Zhang, Y., Yang, T., ... & Xu, Z. (2020). The neuro-immune microenvironment of acupoints—initiation of acupuncture effectiveness. Journal of Leucocyte Biology, 108(1), 189-198.

[2] Zhu, J., Li, J., Yang, L., & Liu, S. (2021). Acupuncture, from the ancient to the current. The anatomical record, 304(11), 2365-2371.

[3] Xu, Y., Guo, Y., Song, Y., Zhang, K., Zhang, Y., Li, Q., ... & Guo, Y. (2018). A new theory for acupuncture: promoting robust regulation. Journal of Acupuncture and Meridian Studies, 11(1), 39-43.

[4] Wang, S., Fang, R., Huang, L., Zhou, L., Liu, H., Cai, M., ... & Akkaif, M. A. (2024). Acupuncture in traditional Chinese medicine: a complementary approach for cardiovascular health. Journal of Multidisciplinary Healthcare, 3459-3473.

[5] Chen, D., Yang, G., & Zhou, K. (2015). Traditional theories and the development of motion acupuncture: A historical perspective. Int J Clin Acupunct, 24(4), 223-227.

[6] Qi, W., He, B., Gu, Q., Li, Y., & Liang, F. (2024). Scientific exploration and hypotheses concerning the meridian system in traditional Chinese medicine. Acupuncture and Herbal Medicine, 4(3), 283-289.

[7] Li, F., He, T., Xu, Q., Lin, L. T., Li, H., Liu, Y., ... & Liu, C. Z. (2015). What is the Acupoint? A preliminary review of Acupoints. Pain Medicine, 16(10), 1905-1915.

[8] Li, N., Guo, Y., Gong, Y., Zhang, Y., Fan, W., Yao, K., ... & Lyu, Z. (2021). The anti-inflammatory actions and mechanisms of acupuncture from acupoint to target organs via neuro-immune regulation. Journal of inflammation research, 7191-7224.

[9] Lin, J. G., Kotha, P., & Chen, Y. H. (2022). Understandings of acupuncture application and mechanisms. American journal of translational research, 14(3), 1469.

[10] Zhang, Y. (2021). Interpretation of acupoint location in traditional Chinese medicine teaching: Implications for acupuncture in research and clinical practice. The Anatomical Record, 304(11), 2372-2380.

[11] Godson, D. R., & Wardle, J. L. (2019). Accuracy and precision in acupuncture point location: a critical systematic review. Journal of acupuncture and meridian studies, 12(2), 52-66.

[12] Lee, Y. S., Ryu, Y., Yoon, D. E., Kim, C. H., Hong, G., Hwang, Y. C., & Chae, Y. (2020). Commonality and specificity of acupuncture point selections. Evidence-Based Complementary and Alternative Medicine, 2020(1), 2948292.

[13] Han, X., Li, X., Liang, Y., Wang, X., Xu, D., & Guan, R. (2021, December). Acupuncture and tuina knowledge graph for ancient literature of traditional chinese medicine. In 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (pp. 674-677). IEEE.

[14] Zhou, B. Y., Zhu, C. F., Li, M., Zhang, N., Jia, Y. M., & Wu, A. Q. (2023). Study on the explicitation of implicit knowledge and the construction of knowledge graph on moxibustion in medical case records of ZHOU Mei-sheng's Jiusheng. Chinese Acupuncture & Moxibustion, 584-590.

[15] Zheng, H. D., Wang, Z. Q., Li, S. S., Lu, Y., Huang, Y., Zhou, C. L., ... & Wu, H. G. (2020). Effect of acupoints on acupuncture-moxibustion and its therapeutic mechanism. World Journal of Traditional Chinese Medicine, 6(3), 239-248.

[16] Zhao, Y., Huang, L., Liu, M., Gao, H., & Li, W. (2021). Scientific knowledge graph of acupuncture for migraine: a bibliometric analysis from 2000 to 2019. Journal of Pain Research, 1985-2000.

[17] Li, X., Han, X., Wei, S., Liang, Y., & Guan, R. (2024). Acupuncture and tuina knowledge graph with prompt learning. Frontiers in big Data, 7, 1346958.

[18] Yanqing Zhao,Li Huang & Wentao Li. (2024). Mapping knowledge domain of acupuncture for Parkinson's disease: a bibliometric and visual analysis. Frontiers in Aging Neuroscience1388290-1388290.

[19] Xiaohui Cui,Yu Yang,Dongmei Li,Xiaolong Qu,Lei Yao,Sisi Luo & Chao Song. (2023). Fusion of SoftLexicon and RoBERTa for Purpose-Driven Electronic Medical Record Named Entity Recognition. Applied Sciences(24).

[20] Aashita Chhabra & Mansaf Alam. (2025). RegGRU-Opt: A Robust GRU-Based RNN Model for High-Performance Sentiment Analysis and Binary Classification. Arabian Journal for Science and Engineering(prepublish),1-40.

[21] Haowei Yang,Longfei Yun,Jinghan Cao,Qingyi Lu & Yuming Tu. (2024). Optimization and Scalability of Collaborative Filtering Algorithms in Large Language Models. Academic Journal of Computing & Information Science(12).

[22] Kazım Hanbay& Taha Burak Özdemir. (2024). Ship Classification Based On Co-Occurrence Matrix and Support Vector Machines. ELECTRICA(3),812-817.