Chen *et al. BMC Medical Informatics and Decision Making* 2019, **19**(Suppl 6):264
https://doi.org/10.1186/s12911-019-0963-0

BMC Medical Informatics and Decision Making

RESEARCH Open Access

Heterogeneous information network based clustering for precision traditional Chinese medicine

Xintian Chen[1,3†], Chunyang Ruan[1,3†], Yanchun Zhang[2,3*] and Huijuan Chen[4]

From IEEE International Conference on Bioinformatics and Biomedicine 2018
Madrid, Spain. 3-6 December 2018

Abstract

Background: Traditional Chinese medicine (TCM) is a highly important complement to modern medicine and is widely practiced in China and in many other countries. The work of Chinese medicine is subject to the two factors of the inheritance and development of clinical experience of famous Chinese medicine practitioners and the difficulty in improving the service capacity of basic Chinese medicine practitioners. Heterogeneous information networks (HINs) are a kind of graphical model for integrating and modeling real-world information. Through HINs, we can integrate and model the large-scale heterogeneous TCM data into structured graph data and use this as a basis for analysis.

Methods: Mining categorizations from TCM data is an important task for precision medicine. In this paper, we propose a novel structured learning model to solve the problem of formula regularity, a pivotal task in prescription optimization. We integrate clustering with ranking in a heterogeneous information network.

Results: The results from experiments on the Pharmacopoeia of the People's Republic of China (ChP) demonstrate the effectiveness and accuracy of the proposed model for discovering useful categorizations of formulas.

Conclusions: We use heterogeneous information networks to model TCM data and propose a TCM-HIN. Combining the heterogeneous graph with the probability graph, we proposed the TCM-Clus algorithm, which combines clustering with ranking and classifies traditional Chinese medicine prescriptions. The results of the categorizations can help Chinese medicine practitioners to make clinical decision.

Keywords: TCM, Formula, Heterogeneous Information network, Clustering, Ranking

Background

Traditional Chinese medicine(TCM) has a long history and is one of the oldest forms of medicine. The fact that traditional Chinese medicine can exist for thousands of years is a proof that TCM has the value of its medical form. Traditional Chinese medicine is being accepted by the public. More and more researchers are working on Chinese medicine, and more Chinese medicines are used in different countries.[1].

The same disease may have different symptoms in different patients. Due to the differences between patients, accurate diagnosis and treatment are even more important[2]. In the field of Western medicine, doctors explore the cause of the disease and focus on treating specific parts of the body. However, Chinese medicine works differently. Traditional Chinese medicine and Western medicine have fundamental differences in diagnosis and treatment. The Chinese medicine practitioner explores the internal and external causes of the patient and combines them accordingly.

In TCM, herbal remedies are usually based on traditional Chinese medicine formula, and the use of only one type of herbal medicine rarely occurs. Each herb has its

*Correspondence: yanchunzhang@fudan.edu.cn
†Xintian Chen and Chunyang Ruan contributed equally to this work.
[2]College of Engineering and Science, Victoria University, Melbourne, Australia
[3]Cyberspace Institute of Advanced Technology, Guangzhou University, Guangzhou, China
Full list of author information is available at the end of the article

advantages and disadvantages, and they are formulated in a reasonable proportion. Figure 1 is a schematic diagram of the composition of a formula [20]. Formula is the foundation of traditional Chinese medicine and the core of TCM research. Traditional Chinese medicine data such as ancient Chinese literature and clinical prescriptions contain prescription data. How to study prescription data through scientific and technical analysis has become the main topic of TCM informatization. Traditional Chinese medicine data generally appears in the form of texts with strong natural language, usually characterized by unstructured, massive and heterogeneous. These characteristics have become a huge challenge in the process of informationization of Chinese medicine. In order to meet the challenge, how to integrate the data of heterogeneous Chinese medicine and model representation, using the data processing method to analyze Chinese medicine data has become an important work of TCM informationization.[3].

The traditional model can no longer meet the inheritance and development of TCM knowledge. The main bottleneck is the organization form of the knowledge and the limitation of human resources. Simply, TCM knowledge mainly exists in the forms of prescriptions, medical records, etc. Unlike ordinary texts, TCM texts have irregular natural language, which creates great difficulties for digitization. At the same time, because the degree of informationization of TCM is not high, the inheritance model of TCM is generally an apprenticeship mode in which an old practitioner cultivates apprentices. As a result, some knowledge cannot be passed down in time. How to pass on the vast knowledge of TCM in an efficient way has become a hot topic in the field of TCM research.

The development of machine learning and artificial intelligence is conducive to drive the inheritance of traditional Chinese medicine. Experience can be regarded as a kind of knowledge in artificial intelligence, which can be used as input data for machine learning. Secondly, "dialectical treatment" is the basic principle of TCM diagnosis of diseases. This is in line with the basic principles of machine learning: the model is trained based on the training set, and the model gives the target value based on the input values. Xu et al. used data mining methods to explore drug combinations for nonalcoholic fatty liver disease[4]. Chen et al. used a three-part map to explore the symptom-disease pattern in the case[5]. Liu et al. used CRF to learn the characteristic patterns in TCM cases to identify symptoms and cases[6]. Wang et al. proposed a probabilistic model for the analysis of symptoms, diseases, and drug relationships in TCM cases[7]. Although some of these tasks can also learn the low-dimensional representation of nodes, they focused on data with rich semantics. The majority of TCM medical data, particularly formula-based prescriptions, are lack of good semantic information.

In this paper, we propose a clustering algorithm based on probability model to solve the clustering problem of heterogeneous information network of traditional Chinese medicine. For a given target type, we aim to generate the clustering of the target object and the ranking information of the objects in the cluster. We propose a heterogeneous information network of traditional Chinese medicine, which is a star network schema. The algorithm can obtain stable clustering results after many iterations. Our main contributions are as follows:

We propose a clustering algorithm based on probability model, which integrates clustering with ranking information for Chinese medicine formula categorization and discover potential knowledge. The algorithm can help doctors optimize diagnosis and prescription. According to the ranking information of each object in the cluster, doctors can easily assess its importance.

We conducted experiments on real data sets of traditional Chinese medicine. The experimental results show

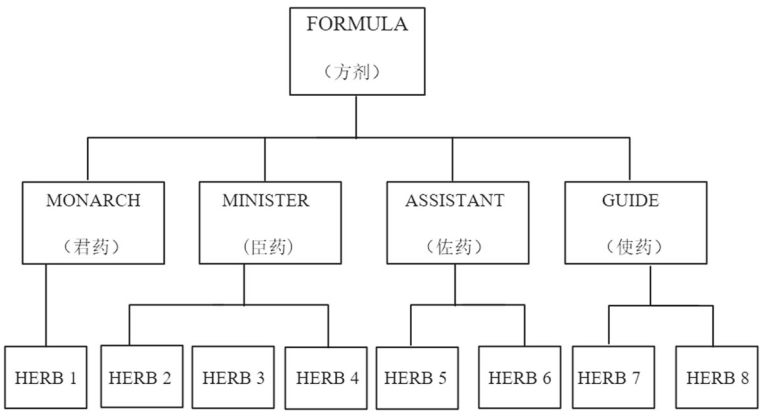

Fig. 1 The composition of a formula

that the algorithm is effective and accurate. The algorithm can provide reasonable clustering results for optimizing prescriptions and is confirmed by Chinese medicine experts.

Social networks, the Internet, medical information networks and many other networks in real world contain a large number of interconnected nodes. These networks are called information networks [8]. The ubiquitous information network is an important part of modern information infrastructure. The nodes in the information network are connected by an intricate network structure, which contains rich information. At present, information network analysis is not only widely concerned by researchers in various fields, but also a hot topic in the field of data mining and information retrieval. However, most information network related research has a basic assumption: the types of nodes and the types of links in the network are unique. That is to say, the researcher does not distinguish the types of nodes and regards them as homogeneous information networks, for example, the author collaboration network. In fact, these networks are full of different kinds of nodes, and it is more reasonable to think of them as heterogeneous information networks (HINs) with different types of nodes and links. Heterogeneous information networks contain richer semantic information in nodes and links. For example, in a bibliographic information network, papers are connected to each other by different types of nodes, such as authors, conferences, and topics. If a paper is connected to two authors at the same time, the two authors have a cohesive relationship with this paper[9].

Ranking is an important task on the heterogeneous information network, and it faces some challenges. First, there are different types of objects and relationships in HIN. Second, different types of objects and relationships have different semantic information. In addition, the ranking information of different objects will affect each other.Taking the bibliographic heterogeneous network as an example, ranking on authors may have different results under different meta paths [10] since these meta paths will construct different link structures among authors. Moreover, the rankings of different-typed objects have mutual effects. For example, reputable authors generally publish papers in top journals[11].

Clustering is a process of classifying similar objects. The objects in the same cluster are similar, and the objects between different clusters are dissimilar. Traditional clustering is generally based on object-based features, such as the K-means algorithm. At present, network-based clustering (community discovery) and other issues are receiving widespread attention. The correlation model usually treats it as a homogeneous information network and divides the network into a series of subgraphs in a given way (e.g., normalized cuts and modularity). Many algorithms have been proposed to solve this NP-hard problem, such as the spectral method [12], greedy method and sampling technique [13]. Some studies consider both the link information and attribute information of the object to improve clustering accuracy [14]. Further, clustering on heterogeneous information networks has received attention.

Unlike homogeneous networks, different types of objects on heterogeneous information networks present a huge challenge to the task.

On the one hand, different types of objects in the network bring new forms of clustering. For example, a cluster may contain different types of objects with the same topic. A cluster of database domains contains authors, conferences, and papers in this field. In this case, clustering on heterogeneous information networks has richer semantics, but it also faces more challenges. On the other hand, the rich information contained in the network helps to improve the accuracy of the task. Li et al. put forward the SCHAN algorithm to solve the clustering problem in Attributed HIN[15]. Zhou et al. designed a dynamic learning algorithm SI-Cluster for social influence based graph clustering[16]. Luo et al. introduced the concept of relation-path to measure the similarity between two objects and propose a framework for semi-supervised learning in HINs[17]. Undoubtedly, these approaches improved the clustering performance, but they were confined to entities with rich attributes or labeled data.

Methods

Problem formulation

In this section, we introduce several important concepts and define the problem of clustering in the TCM HIN.

Definition 1 (Heterogeneous Information Network)**.** An information network is defined as an undirected graph $G =< V, E >$ with an object type mapping function $\tau : V \rightarrow T$ and link type mapping function $\psi : E \rightarrow R$, where $T = \{T_k\}_{k=1}^{|T|}$ is a set of object types and $R = \{R_k\}_{k=1}^{|R|}$ is a set of link types on T. Specifically, we call such an information network a HIN when $|T| \geq 2$ and a homogeneous information network when $|T| = 1$.

Definition 2 (Network schema)**.** Given a HIN $G =< V, E >$, a network schema is defined as an undirected graph $S_G =< T, R >$, where $T = \{T_k\}_{k=1}^{|T|}$ is a set of object types and $R = \{R_k\}_{k=1}^{|R|}$ is a set of link types on T.

Definition 3 (Star Network)**.** A HIN $G =< V, E >$ on $|T|$ types of entities $T = \{T_k\}_{k=0}^{K} (K \geq 2)$ is with a star network schema if, $\forall e =< t_i, t_j > \in E, t_i \in T_0 \wedge t_j \in T_k (k \neq 0)$, or vise versa. G is then called a star network. T_0 is called the target type, and $T_k (k \neq 0)$ are called attribute types[11].The schema for TCM-HIN is shown in Fig. 2.

In this paper, we use T to represent the set of types of TCM entities. We have $T = \{F_m, F_c, H, S\}$, where F_m, F_s, H, and S denote the entity types "formula",

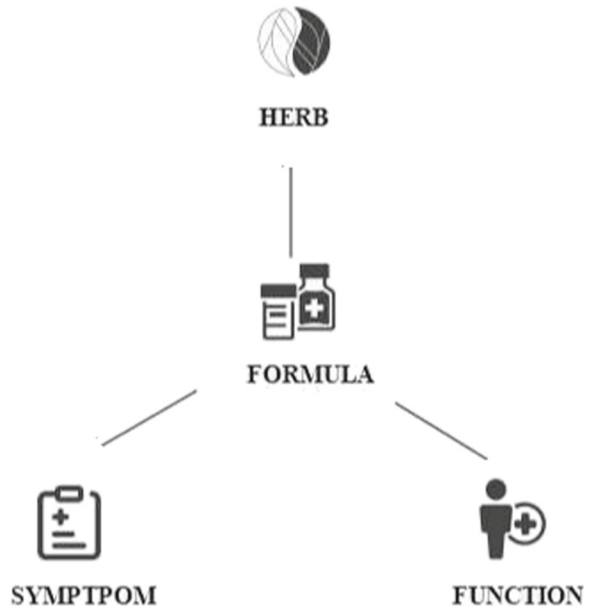

Fig. 2 Schema for TCM-HIN

"function", "herb", and "symptom", respectively. For convenience, we use F_m to denote both the set of objects belonging to the "formula" type and the type name. Other types are similar to F_m. We use $R = \{F_mF_c, F_mH, F_mS\}$ to represent the set of types of TCM relations on T, where F_mF_c, F_mH, and F_mS denote the relation types "formula-function", "formula-herb", and "formula-symptom", respectively.

Definition 4 (TCM-HIN). TCM-HIN is a HIN $G =< V, E >$ with star network schema $S_G = (T, R)$, where $T = \{F_m, F_c, H, S\}$ and $R = \{F_mF_c, F_mH, F_mS\}$[18]. An example of TCM-HIN is shown in Fig. 3.

Based on these definitions, we can formulate our key problem as follows: given a TCM-HIN $G =< V, E >$, the target type T_0, and a specified cluster number K, we aim to generate K clusters $\{C_K\}$ for target objects from target type on G, as well as the within-cluster ranking information for all the objects based on these clusters in the network.

We propose a ranking-based clustering algorithm for mining formula categorization. In this section, we first introduce the overall clustering framework. Then, we explain four important parts of the algorithm in detail.

Framework of algorithm

To integrate ranking with clustering in a HIN, a model is required to flexibly support these two tasks. Therefore, we propose a probabilistic generative model to estimate the probability of target and attribute objects in the network. We can use the rankings of objects to infer the probability of objects and clustering information. The major difficulty in clustering in a HIN is the definition and calculation of pairwise similarity between objects. We map each target object into a low-dimensional space defined by the current clustering result to avoid defining and calculating similarity between each pair of objects.

TCM clustering is mainly composed of the following five steps:

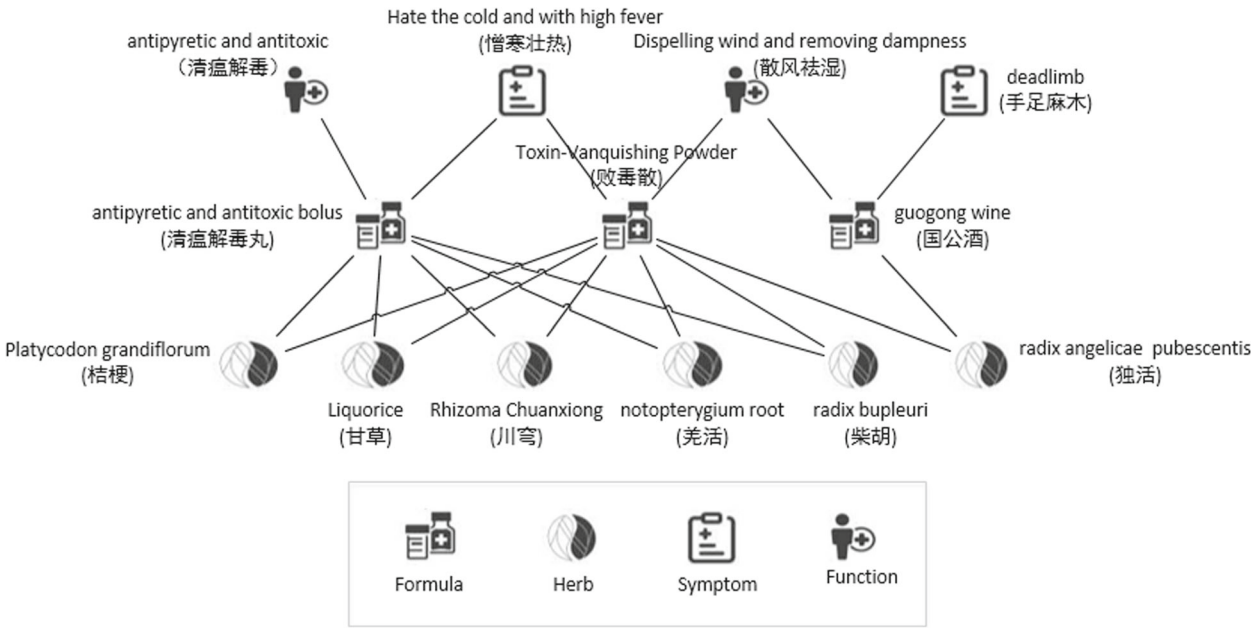

Fig. 3 An example of TCM-HIN

• Step 0: Randomly initialize partitions of target objects and induce initial clusters from the original network according to these partitions, i.e., $\{C_k^0\}_{k=1}^K$. Decompose the star schema network into three bipartite networks, where $V = \{F_m, S\}, \{F_m, H\}$, and $\{F_m, S\}$, respectively.

• Step 1: For each bipartite network, build a ranking-based probabilistic generative model for target type and attribute type, i.e., $\{P(x|C_k^t)\}_{k=1}^K$.

• Step 2: For each bipartite network, estimate the posterior probabilities to each cluster for each target object, i.e., $\{P(C_k^t|x)\}_{k=1}^K$.

• Step 3: Calculate the distance from each target object to each cluster center based on the posterior probabilities and then assign each target object to the nearest cluster.

• Step 4: Repeat Steps 1, 2 and 3 until the cluster does not change significantly or the iteration number is larger than a predefined number.

Algorithm 1 TCM-Clus

Require: Cluster number K and relation matrix W
Ensure: K clusters: $\{C_k\}_{k=1}^K$

```
induce initial clusters {C_k^0}_{k=1}^K from random partitions of target objects
decompose star schema network into three bipartite networks
while nonconvergence do
    for each bipartite network do
        build ranking-based probabilistic generative model: {P(x|C_k^t)}_{k=1}^K
        estimate the posterior probabilities: {P(C_k^t|x)}_{k=1}^K
    end for
    calculate the distance and then assign target object to the nearest cluster
end while
```

The core framework of TCM-Clus is shown in Algorithm 1. In TCM-HIN, a formula may connect to more than one herb, function, and symptom. For example, in Fig. 4, a formula called 桂星降脂汤 contains two herbs called 肉桂(Cinnamomum cassia) and 制南星(arisacma consanguineum) and has two functions called 化痰祛风(dispelling pathogenic wind and eliminating phlegm) and 益火消阴(boosting source of fire for eliminating abundance of yin). However, it does not mean that 肉桂(Cinnamomum cassia) has both functions. Therefore, we should decompose the TCM-HIN into several bipartite networks as above, instead of simply making estimation in original TCM-HIN[11]. In this paper, we decompose the TCM-HIN into three bipartite networks(G_S, G_H, G_{F_c}), which are induced graphs of the original graph G. Because the ranking function and posterior probability estimation for each bipartite network are similar, we only present the explanation for the bipartite network $G_S =< V, E >$, where $V = \{F_m, S\}$ and $G_S \subseteq G$.

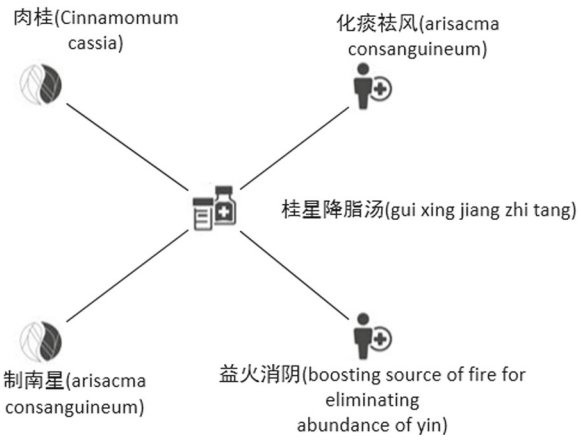

Fig. 4 Schema for TCM-HIN

Ranking function

In information network analysis, the two most important ranking algorithms are PageRank [19] and HITS [20], both of which are successfully applied to Internet searches. PageRank is a link analysis algorithm that assigns a numerical weight to each object of the information network, with the purpose of "measuring" its relative importance within the object set. Conversely, HITS ranks objects based on two scores: authority and hub. Authority estimates the value of the content of the object, whereas hub measures the value of its links to other objects. Both PageRank and HITS evaluate the static quality of objects in the information network, which is similar to the intrinsic meaning of our ranking methods. However, both PageRank and HITS are designed on a network of webpages, which is a directed homogeneous network, and the weight of the edge is binary.

Definition 5 (Ranking Distribution and Ranking Function)**.** A ranking distribution $P(\mathbf{T})$ on a type of object T is a discrete probability distribution, which satisfies $P(T = t) \geq 0(\forall t \in T)$ and $\sum_{t \in T} P(T = t) = 1$. A function $f : G \rightarrow P(\mathbf{T})$ defined on an information network G is called a ranking function on type T if given an information network G, it can output a ranking distribution $P(\mathbf{T})$ on T.

Ranking is beneficial for people to grasp the importance of objects in a collection. For example, PageRank and authority of HITS represent the static importance of webpages, while the rank of a document to a given query in text retrieval reflects the relevance of the document to that query.

We use W to represent the adjacency matrix, which we call the relation matrix, between the target type and the

attribute type. We can define the matrix as

$$W_{F_mS}(i,j) = p_{ij}$$

where i and j are two objects from type F_m and type S and p_{ij} is the frequency of i that links to j.

We have two simple empirical rules:

• Rule 1: Highly ranked formulas can cure highly ranked symptoms.

• Rule 2: One highly ranked symptom can enhance the rank of another symptom if they are cured by the same formula.

According to Rule 1, we generate the ranks of types F_m and S as follows:

$$P(f_{mi}|F_m, G) = \sum_{j=1}^{|S|} W_{F_mS}(f_{mi}, j)P(s_j|S, G) \quad (1)$$

$$P(s_j|S, G) = \sum_{i=1}^{|F_m|} W_{SF_m}(s_j, i)P(f_{mi}|F_m, G) \quad (2)$$

where G is a network, f_{mi} is an object from type F_M, and s_j is an object from type S. Notice that the normalization will not change the ranking position of an object, but it provides a relative importance score to each object. After normalization, we have

$$P(\mathbf{F_m}|F_m, G) = \frac{W_{F_mS}P(\mathbf{S}|S, G)}{\|W_{F_mS}P(\mathbf{S}|S, G)\|} \quad (3)$$

$$P(\mathbf{S}|S, G) = \frac{W_{SF_m}P(\mathbf{F_m}|F_m, G)}{\|W_{SF_m}P(\mathbf{F_m}|F_m, G)\|} \quad (4)$$

We can prove that $P(\mathbf{F_m}|F_m, G)$ is the eigenvector of $W_{F_mS}W_{SF_m}$ and $P(\mathbf{S}|S, G)$ is the eigenvector of $W_{SF_m}W_{F_mS}$.

Proof Combining (3) and (4), we can obtain

$$\begin{aligned} P(\mathbf{F_m}|F_m, G) &= \frac{W_{F_mS}P(\mathbf{S}|S, G)}{\|W_{F_mS}P(\mathbf{S}|S, G)\|} \\ &= \frac{W_{F_mS}\frac{W_{SF_m}P(\mathbf{F_m}|F_m, G)}{\|W_{SF_m}P(\mathbf{F_m}|F_m, G)\|}}{\|W_{F_mS}\frac{W_{SF_m}P(\mathbf{F_m}|F_m, G)}{\|W_{SF_m}P(\mathbf{F_m}|F_m, G)\|}\|} \\ &= \frac{W_{F_mS}W_{SF_m}P(\mathbf{F_m}|F_m, G)}{\|W_{F_mS}W_{SF_m}P(\mathbf{F_m}|F_m, G)\|} \end{aligned}$$

Thus, $P(\mathbf{F_m}|F_m, G)$ is the eigenvector of $W_{F_mS}W_{SF_m}$. Similarly, $P(\mathbf{S}|S, G)$ is the eigenvector of $W_{SF_m}W_{F_mS}$. We can use the power method to calculate the primary eigenvector. □

When considering Rule 2, we can revise the equation as

$$\begin{aligned} P(s_j|S, G) = \alpha &\sum_{i=1}^{|F_m|} W_{SF_m}(s_j, i)P(f_{mi}|F_m, G) \\ &+ (1-\alpha)\sum_{i=1}^{|S|} W_{SS}(j, i)P(s_j|S, G) \end{aligned} \quad (5)$$

where $W_{SS} = W_{SF_m}W_{F_mS}$ and parameter $\alpha \in [0, 1]$ determines the weight of "symptom-formula" and "symptom-symptom". Similarly, we can prove that $P(\mathbf{S}|S, G)$ should be the primary eigenvector of $\alpha W_{SF_m}W_{F_mS} + (1-\alpha)W_{SS}$, and $P(\mathbf{F_m}|F_m, G)$ should be the primary eigenvector of $\alpha W_{F_mS}(I - (1-\alpha)W_{SS})^{-1}W_{SF_m}$.

In fact, if we consider the problem from the perspective of the meta path, these two rules reflect the meta path based relationship between objects. Rule 1 corresponds to meta path $S - F_m$, while Rule 2 corresponds to meta path $S - F_m - S$.

Ranking-based probabilistic generative model

We assume that the probabilities that objects from different types will be visited in the given network are independent of each other. The probability of visiting an object in G can be decomposed into two parts:

$$p(x|G) = p(T_x|G) \times p(x|T_x, G)$$

where the first part $p(T_x|G)$ is the general probability that the type of x will be visited in the network G and the second part $p(x|T_x, G)$ is the probability that an object x will be visited among all the objects from type T_x in the network G. Here, we consider the ranking distribution as the probability of objects to be visited within their own type in a given information network G. We will show that the value of $p(T_x|G)$ is not important and can be set to 1 later. In a subnetwork $G_k = G(C_k)$, we can calculate the probability of visiting an object:

$$p(x|G_k) = p(T_x|G_k) \times p(x|T_x, G_k)$$

However, we will encounter problems if we use the above equation directly. In a given cluster, a target object may link to objects whose ranking is zero in that cluster. In addition, a target object may not belong to the current cluster. If we simply assign the probability of visiting the target object as zero in that cluster, then we will lose some important information. To solve this problem, we can use smoothing, which is a well-known technique in information retrieval to cope with the zero probability problem for missing terms in a document [21]. We add the global ranking to smooth the conditional ranking before calculating the visibility for the target object:

$$p(x|T_x, G_k) = (1-\lambda)p(x|T_x, G_k) + \lambda p(x|T_x, G) \quad (6)$$

where the smoothing parameter λ denotes the portion of global ranking.

To evaluate the model, we make another independence assumption that the probabilities that objects from the same types will be visited are also independent of each other:

$$p(x_i, x_j|T_x, G) = p(x_i|T_x, G) \times p(x_j|T_x, G) \quad (7)$$

where $x_i, x_j \in T_x$.

Posterior probability estimation using EM algorithm

To determine which cluster target objects belong to, we estimate the posterior probability for each target object. For convenience, we use X and Y to represent types F_m and S, where $|X| = m$ and $|Y| = n$.

Given a clustering on the input network G, we can calculate the posterior probability for each target object using the Bayesian rule:

$$p(G_k|x_i) \propto p(x_i|G_k) \times p(k)$$

, where $p(x_i|G_k)$ is the probability that target object x_i will be visited in cluster k and $p(k)$ denotes the relative size of cluster k. From this formula, we can see that type probability $p(T|G)$ is just a constant for calculating the posterior probabilities for target objects and can be neglected.

Let Θ be the parameter matrix, which is an $m \times K$ matrix: $\Theta_{m\times k} = \{P(G_k|x_i)\}(i = 1, 2, \cdots, m; k = 1, 2, \cdots, K)$. To obtain the best Θ that maximizes the likelihood to generate the whole bipartite network, we have the following likelihood function:

$$L(\Theta|W_{XY}) = P(W_{XY}|\Theta) = \prod_{i=1}^{m}\prod_{j=1}^{n} P(x_i, y_j|\Theta)^{W_{XY}(i,j)}$$

, where $P(x_i, y_j|\Theta)$ is the probability of generating link $< x_i, y_j >$ given the current parameter. Because it is difficult to maximize L directly, we apply the EM algorithm to solve the problem. In the E-Step, we introduce hidden variable $z \in \{1, 2, \cdots, K\}$ to represent the cluster label that a link $< x, y >$ is from. The complete log likelihood can be written as

$$\begin{aligned}\log L &= \log \prod_{i=1}^{m}\prod_{j=1}^{n} P(x_i, y_j, z|\Theta)^{W_{XY}(i,j)} \\ &= \log \prod_{i=1}^{m}\prod_{j=1}^{n} [\, p(x_i, y_j|z, \Theta)p(z|\Theta)]^{W_{XY}(i,j)} \\ &= \sum_{i=1}^{m}\sum_{j=1}^{n} W_{XY}(i,j) \log(p(x_i, y_j|z)p(z|\Theta))\end{aligned}$$

Initially, we can set the parameters in $\Theta^{(0)}$ as even values. The expectation of the log likelihood under the current distribution of Z is

$$\begin{aligned}&Q(\Theta, \Theta^{(t)}) \\ &= \sum_{k=1}^{K}\sum_{i=1}^{m}\sum_{j=1}^{n} [\, W_{XY}(i,j) \\ &\quad \times \log(P(x_i, y_j|z = k)P(z = k|\Theta^{(t)}))P(z = k|x_i, y_j, \Theta^{(t)})] \\ &= \sum_{i=1}^{m}\sum_{k=1}^{K}\sum_{j=1}^{n} [\, W_{XY}(i,j) \log(P(z = k|\Theta^{(t)}))P(z = k|x_i, y_j, \Theta^{(t)})] \\ &\quad + \sum_{k=1}^{K}\sum_{i=1}^{m}\sum_{j=1}^{n} [\, W_{XY}(i,j) \log(P(x_i, y_j|z = k))P(z = k|x_i, y_j, \Theta^{(t)})]\end{aligned} \tag{8}$$

where $\Theta^{(t)}$ is the parameter matrix after t iterations.

We can use the Bayesian rule to calculate conditional distribution $P(z = k|x_i, y_j, \Theta^{(t)})$ as follows:

$$P(z = k|x_i, y_j, \Theta^{(t)}) \propto p^{(t)}(x_i|k)p^{(t)}(y_j|k)p^{(t)}(z = k) \tag{9}$$

In the M-Step, to obtain $P^{(t+1)}(z = k)$ that maximizes $Q(\Theta, \Theta^{(t)})$, we introduce the Lagrange multiplier λ. For each $P(z = k)$, where $k = 1, 2, \cdots, K$, we have

$$\begin{aligned}&\frac{\partial}{\partial P(z = k)}[\, Q(\Theta, \Theta^{(t)}) + \lambda(\sum_{k=1}^{K} P(z = k) - 1))] = 0 \\ &\Rightarrow \sum_{i=1}^{m}\sum_{j=1}^{n} W_{XY}(i,j)\frac{1}{P(z = k)}P(z = k|x_i, y_j, \Theta^{(t)}) + \lambda = 0\end{aligned}$$

Now, integrating with (9), we can obtain the new estimation for $P(z = k)$:

$$p^{(t+1)}(z = k) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} W_{XY}(i,j)P\left(z = k|x_i, y_j, \Theta^{(t)}\right)}{\sum_{i=1}^{m}\sum_{j=1}^{n} W_{XY}(i,j)} \tag{10}$$

Finally, each parameter in Θ is calculated as

$$P(G_k|x_i) = P(z = k|x_i) = \frac{P(x_i|G_k)P(z = k)}{\sum_{l=1}^{K} P(x_i|G_l)P(z = l)} \tag{11}$$

Cluster assignment

After we obtain the estimations for each target object in each bipartite network, we can represent the target object as a $3K$-dimensional vector

$$\begin{aligned}\vec{s}_{X_i} = (&p_S(G_1|x_i), \ldots, p_S(G_K|x_i), \ldots, \\ &p_{F_c}(G_K|x_i), \ldots, p_H(G_K|x_i))\end{aligned} \tag{12}$$

The centers for each cluster can thus be calculated accordingly, which is the arithmetic mean of $\vec{s}_{X_i}$ for all x_i in each cluster:

$$\vec{s}_{C_k} = \frac{\sum_{x\in C_k} \vec{s}(x)}{|C_k|} \tag{13}$$

where x_i is an object from type F_m and $|X_k|$ is the size of the cluster k.

The distance between an object and cluster is defined by 1 minus cosine similarity:

$$D(x, C_k) = 1 - \frac{\sum_{l=1}^{K} \vec{s}_x(l)\vec{s}_{C_k}(l)}{\sqrt{\sum_{l=1}^{K}(\vec{s}_x(l))^2}\sqrt{\sum_{l=1}^{K}(\vec{s}_{C_k}(l))^2}} \tag{14}$$

Then, we can assign each object to the cluster with the smallest distance.

User-guided clustering

User guidance is critical for clustering objects in the network[22]. Using different types of link information in a

network, different reasonable clustering results can be generated. We take user guidance in the form of object seeds for some clusters as the prior knowledge for the clustering result Θ by modeling the prior as a Dirichlet distribution rather than treating them as hard labeled ones. For each target object x_i, its clustering probability vector $P(\mathbf{G}|x_i)$ is a multinomial distribution, which is generated from some Dirichlet distribution. If x_i is labeled as a seed in cluster k^*, $P(\mathbf{G}|x_i)$ is then modeled as being sampled from a Dirichlet distribution with parameter vector $\lambda_d \mathbf{e}_{k^*} + 1$, where $\mathbf{e}_{k^*}$ is a K-dimensional basis vector, with the k^*th element as 1 and 0 elsewhere. If x_i is not a seed, x_i is then assumed as being sampled from a uniform distribution, which can also be viewed as a Dirichlet distribution with a parameter vector of $\mathbf{1}$ .The density of $P(\mathbf{G}|x_i)$ given such priors is

$$P(\mathbf{G}|x_i, \lambda_d) = \begin{cases} \prod_k P(G_k|x_i)^{\mathbf{1}\{x_i \in G_k\}\lambda_d} & , x_i \textit{is labeled as } k^* \\ 1 & , x_i \textit{ is not labeled.} \end{cases}$$

where $\mathbf{1}\{x_i \in G_k\}$ is an indicator function, which is 1 if $x_i \in G_k$ holds and 0 otherwise. The hyperparameter λ_d is a nonnegative value and controls the strength of users' confidence over the object seeds in each cluster.

Time complexity analysis

The time complexity of TCM-Clus is composed of the following parts. First, the time complexity for ranking is $O(t_1|E|)$, where t_1 is the iteration number and $|E|$ is the number of links. Notice that $|E| \ll |V|^2$ in a sparse network, where $|V|$ is the total number of objects in the network. Second, for the posterior probability estimation, we need to calculate $O(K|E|+K+mK)$ parameters at each iteration, where the time complexity for (9) is $O|K|E||$, the time complexity for (10) is $O(K)$, and the time complexity for (11) is $O(mK)$. Third, the cluster adjustment for each object has complexity $O(mK^2)$. Since we need to compute the distance between each object and each cluster, the dimension of an object is K. In total, the time complexity for TCM-Clus is $O(t_1|E| + t_2(K|E| + K + mK) + mK^2)$, where t_2 is the iteration number of the estimation. If the network if sparse, which is typical in most applications, the time complexity is almost linear to the number of objects in the network.

Results

In this section, we conduct several experiments to show the effectiveness of TCM-Clus. We discuss the evaluation of TCM-Clus. First, we introduce the datasets used in this paper. Then, we discuss the evaluation of TCM-Clus.

Datasets

In this paper, we use the real datasets ChP, The Pharmacopoeia of the People's Republic of China 2015 Edition (http://wp.chp.org.cn/en/index.html), and 3K+ TCM clinical cases mainly in the stomach. We use herb information in Volume I, which contains 2598 types of medicinal materials without classifications, to set up our experiments. ChP is a unstructured corpus and contains various information. We only extract formula, function, herb, and symptom to build TCM-HIN.

Quantitative evaluation

We use FVIC (fraction of vertices identified correctly) to evaluate the clustering accuracy of the clustering results. It has been used in many research projects and is defined as follows:

$$\begin{aligned} &olSet(c, c*) = \{v | v \in c \wedge v \in c^*\} \\ &maxolSet(c, C_K) = max_{c \in C_K}\{|olSet(c, c^*)|\} \\ &FVIC = \sum_{c \in C_F} \frac{maxolSet(c, C_K)}{N} \end{aligned} \quad (15)$$

where C_F and C_K represent the found clusters and known clusters, respectively. c and c^* are clusters in C_F and C_K, respectively. N is the number of objects in the network. FVIC evaluates the average matching degree by comparing each predicted cluster with the most matching real cluster. A higher score indicates a better clustering with respect to the ground truth.

We compare TCM-Clus with spectral clustering, which is the k-way Ncut algorithm and has been used to cluster Western medical records[23]; PaReCat, which has been used to cluster Chinese medical records for the task of patient record categorization[24]; and K-Means, a common clustering technique. In this experiment, we fix the smoothing parameter λ as 0.2 and weight parameter α as 0.8. The accuracy results are shown in Table 1.

We can observe that TCM-Clus achieves the best clustering accuracy on the two datasets. K-Means shows poor performance because our medical data lack semantic information. Spectral has a good result. However, due to omitting the structure of the graph, it has worse performance compared to TCM-Clus. The performance of PaReCat is closest to our algorithm, but it is more suitable for patient record categorization with disease, symptom and herb. We have shown that TCM-Clus can indeed improve clustering accuracy by integrating ranking with clustering.

Table 1 Clustering Accuracy for Two Datasets

	K-Means	PaReCat	spectral	TCM-Clus
Chp	0.473	0.748	0.741	0.825
Clinical cases	0.589	0.835	0.787	0.875

Parameter study

We use clustering accuracy to analyze the effect of different smoothing parameters λ on Chp dataset. We represent three different λs for symptom, herb, function as λ_s, λ_h and λ_f, respectively. We change one type of λ and fix the other two to 0.2. We run TCM-Clus on ChP datasets, and the results are shown in Fig. 5. The results are based on ten different initial partitions. We can observe that TCM-Clus achieves better accuracy when λ is from 0.1 to 0.8. If the smoothing parameter λ is too small or too large, it means that we only consider conditional ranking or global ranking. Too small ($\lambda \to 0$) or too large ($\lambda \to 1$) will decrease the performance of TCM-Clus.

We also examine the impact of iteration number on the clustering accuracy. As shown in Fig. 6, the clustering accuracy is poor when the iteration number is too small. As the iteration number becomes larger, the accuracy improves and then stabilizes.

Lastly, we examine the impact of the weight parameter α and the result is shown in Fig. 7, If the weight parameter α is too small or too large, it means that we only consider one kind of meta path based relationships. Shorter meta paths have more information than longer ones. If $\alpha = 1$, the clustering accuracy equals 0.791, which is larger than $0.765(\alpha = 0)$.

Qualitative evaluation

We apply our methods to investigate whether TCM-Clus can effectively cluster formulas into informative categories. The results are testified by TCM experts, and many of them are widely used in clinical diagnosis. We show the top-10 herbs and formulas in a cluster identified by our method in Table 2 and the top-5 functions and symptoms in a cluster in Table 3.

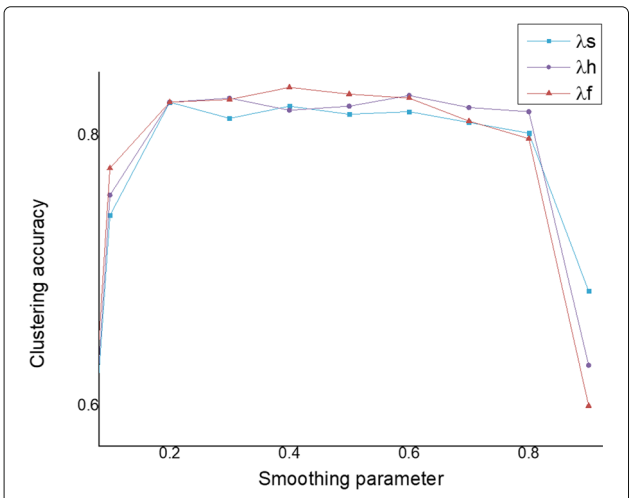

Fig. 5 Clustering accuracy with different smoothing parameters

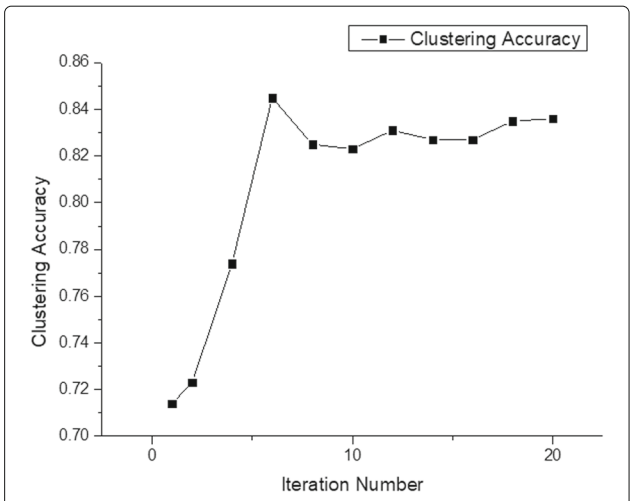

Fig. 6 Clustering accuracy with different iteration numbers

Case evaluation

As mentioned above, TCM-Clus can achieve high quality categorizations. Furthermore, we can obtain new knowledge from clusters, such as "different formulas with similar herbs", "different formulas with similar functions", "different symptoms with similar herbs" and so on. We show an example of "different symptoms with similar herbs" discovered by TCM-Clus in Table 4.

Besides, given a symptom as an input, our system can output proper herb/formula for the symptom. We have listed the herbs used for two symptoms in Table 5. The results are testified by TCM experts, and many of them are widely used in these symptoms.

Discussion

Based on our algorithm, we can learn potential knowledge in TCM, such as discovering similar prescriptions and recommending Chinese medicine based on symptoms. There are still some entities that we have not considered, such

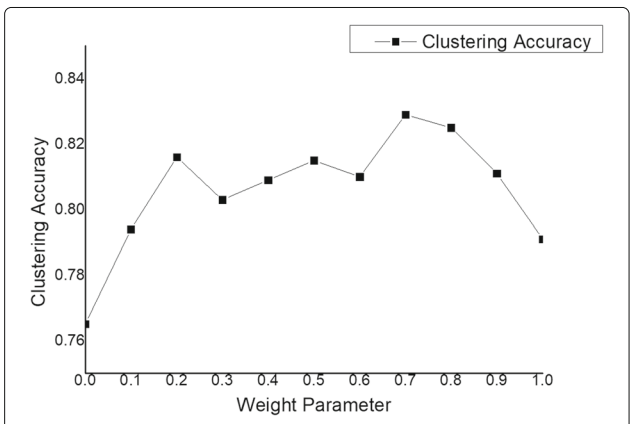

Fig. 7 Clustering accuracy with different weight parameters

Table 2 Top-10 Herbs and Formulas in A Cluster

	Formula	Rank	Herb	Rank
1	天行赤眼方(tian xing chi yan fang)	0.0199	甘草(Chinese liquorice)	0.0612
2	泻肝散	0.0136	连翘(weeping forsythia)	0.0604
3	内疏黄连汤(nei shu huang lian tang)	0.0128	栀子(gardenia)	0.0592
4	疏风清热汤(shu feng qing re tang)	0.0120	黄芪(A. propinquus)	0.0471
5	泻肺饮(xie fei yin)	0.0110	赤芍(Chinese peony root)	0.0425
6	清咽利膈汤(qing yan li ge tang)	0.0104	黄连(Chinese goldthread)	0.0411
7	凉膈散(liang ge san)	0.0102	桔梗(Chinese bellflower)	0.0353
8	还阴救苦汤(huan yin jiu ku tang)	0.0078	当归(female ginseng)	0.0288
9	凉营清气汤(liang ying qing qi tang)	0.0076	石膏(gypsum fibrosum)	0.0279
10	清瘟败毒饮(qing wen bai du yin)	0.0072	荆芥(tenuifolia)	0.0274

Table 3 Top-5 Functions and Symptoms in A Cluster

	Function	Rank	Symptom	Rank
1	清热解毒(clearing heat and detoxifying)	0.0213	目赤肿痛(swelling and pain of eye)	0.0135
2	疏散风热(dispelling wind and heat)	0.0207	面赤唇焦(red face and labial coke)	0.0130
3	清气风热(clearing heat QI)	0.0185	咽痛音哑(sore throat and losing voice)	0.0124
4	凉血解毒(cooling the blood and detoxifying)	0.0143	肺热咳嗽(coughing with lung heat)	0.0117
5	清脏腑热(clearing bowel and visceral heat)	0.0122	湿热痞满(feeling muggy and distension)	0.0106

Table 4 Different Symptoms with Similar Herbs

Symptoms	Herbs	Common Herbs
胃脘痛(epigastric pain), 柴胡(heartburn), 枳壳(belching),白芍(fullness) 烧心(tongue coating),	香附(red thorowax),川穹(bitter orange),陈皮(Chinese peony), 嗳气(Java grass), 痞满(Chuanxiong), 淮山(mandarine peel), 厚朴(magnolia-bark), 麦芽(barley), 柴胡(Chinese liquorice), 苔黄(rice sprout)	甘草(red thorowax), 谷芽(bitter orange), 枳壳(Chinese peony), 食纳欠佳(Java grass), 柴胡(Chuanxiong), 枳壳(mandarine peel), 白芍(Chinese liquorice)
白芍(Poor food and drink), 疲乏无力(Fatigue), 脉弦(stringy pulse), 香附(pink tongue), 川穹(Liver distention and pain)	陈皮(red thorowax),香附(bitter orange),舌淡红(Chinese peony), 当归(Java grass),板蓝根(Chuanxiong),败酱草(mandarine peel), 川穹(female ginseng),肝胀痛(woad root),甘草(Field pennycress), 红枣(Chinese liquorice),陈皮(red dates)	
舌淡(pale tongue), 苔薄(thin coating) 柴胡(Live Qi), 枳壳(stringy pulse),白芍(Lump in breast)	甘草(red thorowax),肝郁气(bitter orange),香附(Chinese peony), 川穹(Java grass),陈皮(Chuanxiong),脉弦(mandarine peel), 三棱(rhizoma sparganii),莪术(curcuma zedoary),生牡蛎(raw oyster), 乳房肿块(Chinese liquorice),甘草(sea-tent),昆布(Sargassum)	

Table 5 An example for our recommendation

Symptom	海藻(pharyngitis)	喉痹(Deficiency of spleen and deficiency of food)
Herb 1	脾虚食少(Picria fel-terrae)	苦玄参(sea-buckthorn)
Herb 2	沙棘(semen oroxyli)	木蝴蝶(fuling)
Herb 3	茯苓(lignum et folium trachelospermi)	络石藤(ginseng)
Herb 4	人参(herba taching)	大青叶(yam)
Herb 5	山药(Lonicera confusa DC.)	山银花(Chinese-date)

as the amount of herb and the information of patients. In our future work, more research is needed to address general HINs with more kinds of entities. In addition, the ranking function is highly related to different domains, and how we can automatically extract rules based on small partial ranking results given by experts could be another interesting problem.

Conclusions

TCM is one of the most important complementary and alternative medicines. However, the complexity and elusiveness of diagnostic methods limit its development and generalization. Formulas are an essential part of TCM. Mining categorizations from TCM medical records is an important task for precision medicine. We present a novel algorithm, TCM-Clus, for mining formula categorization. We use a generative probabilistic model based on ranking to generate the reachable probability of target objects. Meanwhile, Bayesian rules and the EM algorithm are utilized to estimate the posterior probability. The experiments show that TCM-Clus achieves better clustering results than other representative algorithms and is beneficial for enhancing the predictive accuracy of medicine.

Abbreviations

HIN: Heterogeneous information network TCM: Traditional Chinese medicine

Authors' contributions

CR and YZ designed the study. XC and CR designed the model and programmed the algorithm. HC verified the experimental results. All authors were involved in the revision of the manuscript. XC and CR contributed equally.

Funding

This work is supported in part by the Chinese National Natural Science Foundation (No. 61332013), the the Research Innovation Plan of Shanghai Education Commission (No. 14YS026) and the Research Foundation of Shanghai Municipal Health and Family Planning Commission (No. JP013). Publication costs are funded by the Chinese National Natural Science Foundation (No. 61332013)

Availability of data and material

The Chinese Pharmacopoeia 2015 Edition (http://wp.chp.org.cn/en/index.html).

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

About this supplement

This article has been published as part of BMC Medical informatics and Decision Making Volume 19 Supplement 6, 2019: Selected articles from the IEEE BIBM International Conference on Bioinformatics & Biomedicine (BIBM) 2018: medical informatics and decision making. The full contents of the supplement are available online at https://bmcmedinformdecismak.biomedcentral.com/articles/supplements/volume-19-supplement-6.

Author details

[1]School of Computer Science, Fudan University, Shanghai, China. [2]College of Engineering and Science, Victoria University, Melbourne, Australia. [3]Cyberspace Institute of Advanced Technology, Guangzhou University, Guangzhou, China. [4]School of Basic Medical Science, Shanghai University of Traditional Chinese Medicine, Shanghai, China.

Published: 19 December 2019

References

1. Hao P, Fan J, Jing C, Ma L, Yun Z, Zhao Y. Traditional chinese medicine for cardiovascular disease : Evidence and potential mechanisms. J Am Coll Cardiol. 2017;69(24):2952.
2. Zhang Y. Welcome to health information science and systems. Health Inf Sci Syst. 2013;1(1):1.
3. Oti M, Huynen MA, Brunner HG. Phenome connections. Trends Genet. 2008;24(3):103–6.
4. Ming XUH, Wang YM, Zhang TS. Data mining of regularities and rules of compound herbal formulae for nonalcoholic fatty liver disease. Chin J Inf Tradit Chin Med. 2014;21(8):38–41.
5. Chen J, Poon J, Poon SK, Xu L, Sze DMY. Mining symptom-herb patterns from patient records using tripartite graph. Evid Based Complement Alternat Med eCAM. 2015;2015(1):435085.
6. Liu H, Qin X, Fu B. The symptoms and pathogenesis entity recognition of tcm medical records based on crf. In: Ubiquitous Intelligence and Computing and 2015 IEEE Intl Conf on Autonomic and Trusted Computing and 2015 IEEE Intl Conf on Scalable Computing and Communications and ITS Associated Workshops. IEEE; 2016. p. 1479–84. https://doi.org/10.1109/uic-atc-scalcom-cbdcom-iop.2015.267.
7. Wang S, Huang EW, Zhang R, Zhang X, Liu B, Zhou X, Zhai CX. A conditional probabilistic model for joint analysis of symptoms, diseases, and herbs in traditional chinese medicine patient records. In: IEEE International Conference on Bioinformatics and Biomedicine. IEEE; 2017. p. 411–8. https://doi.org/10.1109/bibm.2016.7822553.
8. Shi C, Li Y, Zhang J, Sun Y, Yu PS. A survey of heterogeneous information network analysis. IEEE Trans Knowl Data Eng. 2016;29(1):1.
9. Shi C, Yu P. Heterogeneous information network analysis and applications; 2017. https://doi.org/10.1007/978-3-319-56212-4.
10. Li Y, Shi C, Yu P, Chen Q. HRank: A path based ranking framework in heterogeneous information network. 2014. https://doi.org/10.1007/978-3-319-08010-9_61.
11. Sun Y, Yu Y, Han J. Ranking-based clustering of heterogeneous information networks with star network schema. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, June 28 - July. ACM Press; 2009. p. 797–806. https://doi.org/10.1145/1557019.1557107.
12. Wakita K, Tsurumi T. Finding community structure in mega-scale social networks. In: Proceedings of the 16th international conference on World Wide Web - WWW'07. ACM Press; 2007. p. 1275–6. https://doi.org/10.1145/1242572.1242805.
13. Zhou Y, Cheng H, Yu JX. Graph clustering based on structural/attribute similarities. Proc VLDB Endowment. 2009;2(1):718–29.
14. Yang T, Jin R, Chi Y, Zhu S. Combining link and content for community detection. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM Press; 2009. p. 927–36. https://doi.org/10.1145/1557019.1557120.
15. Li X, Wu Y, Ester M, Kao Cm, Wang X, Zheng Y. Semi-supervised clustering in attributed heterogeneous information networks. In: Proceedings of the 26th International Conference on World Wide Web - WWW '17. ACM Press; 2017. https://doi.org/10.1145/3038912.3052576.
16. Zhou Y, Liu L. Social influence based clustering and optimization over heterogeneous information networks. Acm Trans Knowl Discov Data. 2015;10(1):1–53.
17. Pang W. Semi-supervised clustering on heterogeneous information networks. 2014. https://doi.org/10.1007/978-3-319-06605-9_45.
18. Ruan C, Wang Y, Zhang Y, Ma J, Chen H, Aickelin U, Zhu S, Zhang T. Thcluster: Herb supplements categorization for precision traditional chinese medicine. In: IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE; 2017. p. 417–24. https://doi.org/10.1109/bibm.2017.8217685.
19. Brin S, Page L. The anatomy of a large-scale hyper-textual web search engine. Computer Networks and ISDN Systems, 30,1998.
20. Kleinberg JM. Authoritative sources in a hyperlinked environment. J ACM (JACM). 1999;46(5):604–32.

21. Zhai C, Lafferty J. A study of smoothing methods for language models applied to information re-trieval: ACM Transactions on Information Systems (TOIS; 2004, p. 22.
22. Yin X, Han J, Yu PS. Cross-relational clustering with user's guidance. In: Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Chicago, Illinois, Usa, August; 2005. p. 344–53. https://doi.org/10.1145/1081870.1081910.
23. Bertsimas D, Pandey R, Vempala s, Wang g. Algorithmic prediction of health-care costs. Oper Res. 2008;56(6):1382–92.
24. Huang EW, Wang S, Liu B, Zhou X, Zhai CX. Parecat: Patient record subcategorization for precision traditional chinese medicine. In: ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics. ACM Press; 2016. https://doi.org/10.1145/2975167.2975213.

Publisher's Note