Gong JB, Wang LL, Sun ST *et al.* iBole: A hybrid multi-layer architecture for doctor recommendation in medical social networks. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 30(5): 1073–1081 Sept. 2015. DOI 10.1007/s11390-015-1583-5

iBole: A Hybrid Multi-Layer Architecture for Doctor Recommendation in Medical Social Networks

Ji-Bing Gong (宫继兵), *Member, CCF*, Li-Li Wang* (王立立), *Student Member, CCF*
Sheng-Tao Sun (孙胜涛), *Member, CCF*, and Si-Wei Peng (彭思维)

[1] *School of Information Science and Engineering, Yanshan University, Qinhuangdao 066004, China*
[2] *The Key Laboratory for Computer Virtual Technology and System Integration of Hebei Province, Yanshan University Qinhuangdao 066004, China*

E-mail: {gongjibing, wanglili_ysu, ysusst}@163.com; ysupsw@sohu.com

Received November 17, 2014; revised June 15, 2015.

Abstract In this paper, we try to systematically study how to perform doctor recommendation in medical social networks (MSNs). Specifically, employing a real-world medical dataset as the source in our work, we propose iBole, a novel hybrid multi-layer architecture, to solve this problem. First, we mine doctor-patient relationships/ties via a time-constraint probability factor graph model (TPFG). Second, we extract network features for ranking nodes. Finally, we propose RWR-Model, a doctor recommendation model via the random walk with restart method. Our real-world experiments validate the effectiveness of the proposed methods. Experimental results show that we obtain good accuracy in mining doctor-patient relationships from the network, and the doctor recommendation performance is better than that of the baseline algorithms: traditional Ranking SVM (RSVM) and the individual doctor recommendation model (IDR-Model). The results of our RWR-Model are more reasonable and satisfactory than those of the baseline approaches.

Keywords doctor recommendation architecture, random walk with restart, doctor-patient tie mining, time-constraint probability factor graph model, medical social network

1 Introduction

As the economy develops, people pay more and more attention to the condition of their health. However, due to limited medical resources, most patients have difficulty in finding appropriate doctors to diagnose their issues. Medical social networks (MSNs) play an increasingly important role in people's health care. How to mine and analyze an MSN is a hot research issue that has recently attracted much attention in both industry and research communities. There have been a few studies on social recommendations. However, they almost completely ignore the insufficiency of real medical information and the heterogeneity and diversity of the social relationship. To the best of our knowledge, the whole architecture of doctor recommendations on MSNs has not been explored yet.

In this paper, we try to systematically investigate how to perform doctor recommendation in MSNs. However, as an emerging research topic, several challenges exist in this study.

- The first is how to mine doctor-patient relationships from a real-world medical dataset and extract network features for ranking nodes.
- The second is how to perform doctor recommendation according to network features since traditional information retrieval models, such as the Boolean model[1] and the Vector Space model[2], are limited to computing similarity degree between query keywords and destination doctors.

Regular Paper
Special Section on Social Media Processing
This work was supported by the the National High Technology Research and Development 863 Program of China under Grant No. 2015AA124102, the Hebei Natural Science Foundation of China under Grant No. F2015203280, and the National Natural Science Foundation of China under Grant Nos. 61303130, 61272466, and 61303233.
*Corresponding Author

• The last is how to evaluate the recommendation precision of our method because it is hard to obtain solid results via traditional methods involving subjective processes.

To address these challenges, we undertake the investigation of the doctor recommendation problem with the following approach.

• We propose a method based on a time-constraint probability factor graph model (TPFG)[3] to mine doctor-patient relationships.

• We define and formalize four network features considering doctor recommendation requirements and compute them for ranking nodes.

• We present a novel hybrid multi-layer architecture (namely iBole). In iBole, we propose a doctor recommendation model (namely RWR-Model) via the random walk with restart method (RWR)[4], and evaluate the recommendation precision according to an information retrieval index.

The rest of our paper is organized as follows. Section 2 describes related work. Section 3 gives an overview of our architecture, iBole, and shows how it mines doctor-patient ties via TPFG, extracts network features for ranking nodes, and makes doctor recommendations via RWR. Section 4 describes experimental details and validations of our results, and Section 5 offers concluding remarks.

2 Related Work

Some traditional information retrieval/recommendation models, e.g., the Boolean model[1] and the Vector Space model[2], compute the similarity between query keywords and destination doctors. They consider similarities between query keywords and individual doctors, but ignore relationships between doctors and patients in social networks[5]. Other relevant methods include the low-rank matrix factorization model[6], the content-based method[7], collaborative filtering[8], and a model-based approach[6]. To handle very large datasets, Salakhutdinov *et al.*[9] presented a class of two-layer undirected graphical models, called Restricted Boltzmann Machines. Shen and Jin[8] developed a joint personal and social latent factor (PSLF) model for online social network recommendation. However, our paper focuses on data mining in real healthcare data rather than in traditional online social networks. Gong and Sun[10] proposed IDR-Model, an individual doctor recommendation model via a weighted average method. Both the application background and the operating principle of IDR-Model are different from those of this paper.

A closely related research topic is expertise search, such as expertise search based on candidate vote by Macdonald and Ounis[5], expertise mining from social networks by Tang *et al.*[11-12], and transfer learning from expertise search to Bole search by Yang *et al.*[11] The most basic method to solve the expert matching problem is bipartite graph matching. After obtaining a fully connected weighted bipartite graph, it solves the problem using the classical Hungarian algorithm[13]. More advanced methods include: 1) obtaining keywords by searching on the Internet and then making matches, 2) calculating relevance in order to make matches using the LSI (latent semantic indexing) method[14-15], 3) obtaining an assignment scheme using linear programming[16], and 4) making assignments using the minimum-cost network flow method[17].

Random walk plays an important role in many fields. Tang *et al.*[18] performed a method based on random walk with restart on topic-augmented graphs to calculate relatedness between users. Also, random walk has gained a lot of interest in academic search/recommendation fields[19-20]. Our paper is mainly inspired by recent researches on graph-based learning[19] and semi-supervised learning[21]. Feng and Wang[21] performed random walk with restart for personalized tag recommendation, incorporating heterogeneous information in social tagging systems.

Most recent research focuses on computing the authority degree of objects in the network, and ranking those objects on that basis[11]. The limitations of these methods are: 1) it is difficult to set parameters by users in traditional ranking models (Boolean model, probabilistic model, etc.), and it is hard to detect and avoid model over-fitting and to integrate multiple models; 2) as the evaluation of the authority degree of doctors may be subjective, authority degree scores in the testing data may be biased, making it hard to give a "fair" ranking result. In contrast, our proposed method is more reasonable and effective because we compute the success rate of finding the most appropriate doctor according to both social relationships and network features. In this paper, we design a novel hybrid and multi-layer architecture of doctor recommendation instead of a single model of doctor recommendation. To the best of our knowledge, research about the whole architecture of doctor recommendation in MSNs has not been explored yet.

3 Architecture of iBole

3.1 Problem Statement

The definition of the problem is given below. We are studying the problem of doctor recommendation for patients. The input to the problem includes query keywords w_1, w_2, $\cdots$, w_m and a time-correlated cooperation relationship network $G = (V, E)$, where $V = V^p \cup V^a$ is the node set and E is the edge set. $V^p = \{p_1, p_2, ..., p_{n_p}\}$ denotes the set of disease cases, the diagnosis time of p_i is expressed by t_i, and $V^a = \{a_1, a_2, ..., a_{n_a}\}$ stands for the set of all participants during the treatment. The output is a list of candidate doctors for a patient with disease p_i. Here, query keywords would be those phrases that reflect the needs and characteristics of patients, such as disease type, income level, round-trip distance, and so on. Notations are summarized in Table 1.

Table 1. Definition of Variables

Symbol	Description
G	Time-correlated cooperation tie network
V	Node set
E	Edge set
V^p	Set of disease cases
V^a	Set of all participants during treatment
st_i	Starting time of diagnosis
ed_i	End time of diagnosis
$r(u)$	Ranking score of node u
$w_1, w_2, \cdots, w_m$	Query keywords

3.2 Overview of iBole

Fig.1 shows our hybrid multi-layer architecture for doctor recommendation. It includes the following layers from bottom to top. 1) The first layer is the data source layer, which can also be called the medical social network layer. It is a real disease information dataset including a lot of valid questionnaires from patients. 2) In the second layer, doctor-patient relationships will be mined, and the mining accuracy shall be improved using an optimization procedure, if necessary. 3) In the third layer, four essential features are extracted from the mined relationships, which will be later used for ranking nodes in medical social networks. 4) In the last layer, the proposed recommendation model is performed to obtain a recommendation ranking score to help a patient find the most appropriate doctor.

In this paper, according to iBole, we first mine doctor-patient ties via TPFG, but do not optimize the mined relationships because the dataset is small. Then we define and extract four features from the network: *DomainRel*, *M-index*, *Activity*, and *Uptrend*. At last, we make a doctor recommendation using the proposed RWR-Model. In addition, the data source layer comprises a real disease case dataset with 2 064 pieces of pulse data and 1 330 valid questionnaires from patients. The feedback information includes attitude toward patients, price rationality, diagnosis efficiency, medical technical level, and curative effect. The information is used to compute the satisfaction degree of a patient with regard to his/her doctor.

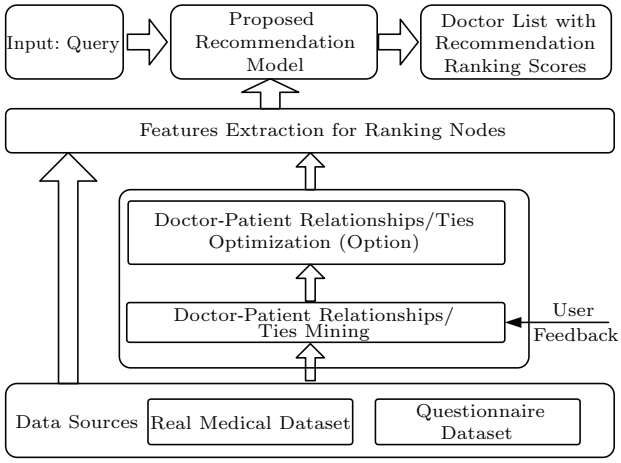

Fig.1. iBole: a hybrid multi-layer architecture for doctor recommendations.

3.3 Mining Doctor-Patient Ties via TPFG

Doctor-patient relationship mining is the basis of accurate doctor recommendations. The task can be formalized as: input a time-correlated cooperation relation network and output a directed acyclic graph. We apply TPFG to mine doctor-patient relationships. In this model, for each patient node a_i, three variables, doctor y_i, the starting diagnosis time st_i, and the end diagnosis time ed_i need to be determined. Given a region feature function $g(y_i, st_i, ed_i)$, to reflect all joint probabilities in the relational graph, we define the joint probability as the product of all region feature functions, as in (1).

$$P(\{y_i, st_i, ed_i\}_{a_i \in V^a}) = \frac{1}{Z} \prod_{a_i \in V^a} g(y_i, st_i, ed_i), \quad (1)$$

where $1/Z$ indicates the normalization factor of the joint probability, with two basic assumptions: 1) a patient knows less about the disease than his/her candidate doctor, and 2) a patient a_i obtains diagnosis results/information later than his/her doctor. In (1), to

obtain the most probable values of all unknown factors, the joint probability needs to be maximized. A large number of unknown parameters would lead to too-large solution space. To reduce the time and space cost performed on the TPFG-based method, we design the rules and corresponding algorithm to filter out those connections that do not stand for doctor-patient cooperation relationships. Thus, we simplify the joint probability problem into the following equation: suppose patient a_i and his/her doctor y_i are determined, and we can obtain $\{st_i, ed_i\} = \arg\max_{st_i<ed_i} g(y_i, st_i, ed_i)$. Then st_i and ed_i can be found.

Before working out this joint probability, we first compute st_i and ed_i, contained in every possible doctor-patient relationship, and then we can obtain a joint probability formula with simplified parameters, as in (2).

$$\begin{aligned} &P(y_1, y_2, \cdots, y_{n_a}) = \frac{1}{z}\prod_{i=1}^{n_a} f_i(y_i|\{y_x|x \in Y_i\}),\\ &f_i(y_i|\{y_x|x \in Y_i\}) = g(y_i, st_{ij}, ed_{ij})\\ &\qquad\qquad \prod_{x\in Y_i} I(y_x \neq i \vee ed_{ij} < st_{xi}),\\ &\quad I(y_x \neq i \vee ed_{ij} < st_{xi})\\ &= \begin{cases} 1, \text{ if } y_x \neq i \vee ed_{ij} < st_{xi},\\ 0, \text{ if } y_x = i \wedge ed_{ij} \geqslant st_{xi}, \end{cases} \end{aligned} \tag{2}$$

where $Y_i = \{y_1, y_2, \cdots, y_{n_a}\} - \{y_i\}$. After simplifying (1), we can use a probability factor graph[22] to solve (2). The factor graph mapped by (2) contains two types of nodes: variable nodes and function nodes. Variable nodes correspond to hidden variables $\{y_i\}_i^{n_a}$. Each variable node y_i links one function node $f_i(y_i|\{y_x|x \in Y_i\})$, which indicates $f_i(y_i|\{y_x|x \in Y_i\})$ is determined by y_i. In addition, the probability factor graph includes one kind of dependence relationship between variables and functions.

3.4 Feature Extraction for Ranking Nodes

To rank nodes and build a random walk with restart model, we design the following four features and extract them from mined doctor-patient relationships.

1) The feature *DomainRel* describes the matching degree between doctors' diagnosis scopes and the disease types of all patients. It is measured from the doctor's point of view and computed using ACT model[23].

2) The feature *M-index* indicates the influence index of a doctor in the healthcare community. Specifically, the *M-index* value is m if one medical technology of a doctor had been used by other doctors at least m times in no more than m disease cases.

3) The feature *Activity* is the activity index of the latest disease cases undertaken by a doctor. It is computed by (3).

$$\begin{aligned} Activity(D) = &\sum_{i=1}^{N} AoT(U_{ty-N+i}(D)) \times\\ &w(ty - N + i), \end{aligned} \tag{3}$$

where $U_{ty-N+i}(D)$ denotes the set of disease cases cured by doctor D in the i-th year in the last N years, $AoT(\cdot)$ means the overall rating scores of the medical effects of cured cases, $w(ty-N+i)$ is the weight value of the $(ty-N+i)$-th year, and N indicates the most recent N years.

4) The feature *Uptrend* describes the uptrend index of a doctor's medical achievements[24]. It is calculated by both (4) and (5).

$$Uptrend(D) = Avg(AoT(U(D))) - C(U(D)) \times Avg(c_i), \tag{4}$$

$$C(D) = \frac{\sum_{i=1}^{N}(c_i \times AoT(U_{ty-N+i}(D))) - N\overline{c} \times \overline{AoT(U(D))}}{\sum_{i=1}^{N}(AoT(AoT(U_{ty-N+i}(D)))^2) - N\overline{AoT(U(D))}^2}, \tag{5}$$

where $U(D)$ denotes the set of disease cases cured by doctor D, $U_{ty-N+i}(D)$ denotes the set of disease cases cured by doctor D in the i-th year during the past N years, $Avg(\cdot)$ represents the average value of all parameters, $C(D)$ indicates a fitted curve that is generated by the least-squares method from all cured cases of doctor D in the last N years, $c_i(= N - i)$ indicates the increment of the number of years from the i-th year to this year during the last N years, $N\overline{c}$ indicates the average value of c_i, U_{ty} means the set of all cured cases in this year, and $\overline{AoT(\cdot)}$ states the average value of the overall rating scores in $AoT(\cdot)$.

We select a classic learning ranking method, Ranking SVM (RSVM)[14-15], as a basic framework for node sorting. To address the problem, RSVM creates a new instance $(x_i^a - x_i^b, z_i)$ for (x_i^a, x_i^b), which is an instance of (y_i^a, y_i^b) having two different ranking levels in query keywords. z_i satisfies: $z_i = +1$ if $y_i^a > y_i^b$; otherwise $z_i = -1$. After building a new training set $\Gamma' = \{(x_i^a - x_i^b, z_i)\}_{i=1}^n$, it is feasible to employ classic RSVM to solve a ranking problem. That is, a sub-optimization problem needs to be solved, as in (6).

$$\min_{\boldsymbol{w}} M\boldsymbol{w} = \frac{1}{2}\|\boldsymbol{w}\| + C\sum_{i=1}^{\ell}\xi_i, \tag{6}$$

where $\xi_i \geqslant 0$, $i = 1, \cdots, \ell, z_i \geqslant 1 - \xi_i$, $\boldsymbol{w}$ denotes a weight vector, ξ_i indicates the ranking error rate of a ranking function, C stands for user-defined parameters of the SVM, every x denotes one instance, (x^a, x^b) is a pair of instances, (x_i^a, x_i^b) is the i-th pair of instance, (x^1, x^2) is the pair of instances which comprises two adjacent instances, y indicates the ranking level of every instance, $(y^a - y^b)$ states the ranking relationship between any two instances, and $(y_i^a - y_i^b)$ refers to the ranking relationship of the i-th pair of instances. The ranking relationship is called Z. $Z = +1$ means the ranking level of x^a is higher than that of x^b, and conversely, $Z = -1$ means the ranking level of x^a is lower than that of x^b. Through training the SVM, weight vectors $\boldsymbol{w}$, which correspond to all feature values in the model shown in (6), can be worked out. Their ranking scores are calculated using this model. Further, we can successfully sort these nodes in our mined medical social network.

3.5 Doctor Recommendation via RWR

Given every participant node u_i, we need to obtain its neighbors to build a medical social network. Given (u_i, v_j) of every directed edge, we define the transition probability from u_i to v_j as in (7).

$$p(u_i, v_j) = \log(\#u_i v_j \pm r(u_i)) \times \log(\#v_j u_i \pm r(v_j)), \quad (7)$$

where $\#u_i v_j$ denotes the number of times that doctor u_i checks patient v_j during the period of diagnosis and treatment, and $\#v_j u_i$ is the number of times that patient v_j visits doctor u_i. $r(u_i)$ and $r(v_j)$ refer to the ranking scores of nodes u_i and v_j, respectively. We can obtain the intimacy transition probability matrix (ITP-Matrix) according to (7), and use ITP-Matrix as the probability matrix of u_i's random walk with restart.

After defining the ITP-Matrix, we can perform the random walk with restart on the medical social network. Thus, we may compute the ranking score of node u_i according to (8) after running every round random walk.

$$RW(u_i) = \alpha \times RW(u_i) + \beta \times \sum\nolimits_{v_j \in R_{u_i}} p(u_i, v_j) \times RW(v_j), \quad (8)$$

where $RW(u_i)$ denotes the ranking score of node u_i, $RW(v_j)$ denotes the ranking score of node v_j, and R_{u_i} denotes the set of all neighbor nodes of u_i on the MSN. $\alpha(= 0.25)$ and $\beta(= 0.75)$ are weighted values. After performing a random walk, every node will have a ranking value that is its recommendation score.

Considering the over-convergence problem in (8), we improve it by introducing the divergence factor $C(u_i, v_j)$, shown in (9). Then the new transition probability can be computed by (10).

$$C(u_i, v_j) = |R_{u_i} \cap R_{v_j}|, \quad (9)$$

$$p_{\text{new}}(u_i, v_j) = \lambda_p p(u_i, v_j) + \gamma \times \frac{C(u_i, v_j)}{C\_MAX}, \quad (10)$$

where $p_{\text{new}}(u_i, v_j)$ denotes the optimal transition probability from u_i to v_j, and $C(u_i, v_j)$ denotes the number of common neighbors (friends) of any two given nodes u_i and v_i. $C\_MAX(= 2\,000)$ is a standardization constant, λ_p and γ are weights, and their values are 0.5.

4 Experiments and Evaluations

4.1 Dataset

It is very difficult to obtain training and testing data for studying the topic of doctor recommendation. But fortunately, we developed PDhms[25], a wearable healthcare monitoring system for human pulse diagnosis. This system helped us collect the real-world medical dataset with 2 064 pieces of pulse data in large-scale clinic experiments performed at the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), in 2009 and 2010, as well as the Hitech Fair of China in Shenzhen in 2010 and 2011. The medical dataset involves many kinds of disease data, doctor information, curing and treatment information, patient personal information, social relationships (e.g., colleague ties), doctor-patient relationships, and patients' evaluations of their doctors. Thus, we can build a real-world medical social network. One patient's personal and diagnostic information from our real medical dataset is illustrated in Table 2.

In addition, we chose 1 330 valid entries to calculate the satisfaction degree of patients (SDP) from 2 064 questionnaires. The feedback information includes attitude toward patients, price rationality, diagnosis efficiency, medical technical level, curative effect, etc. Table 3 shows an instance of questionnaire information to illustrate how to quantitatively reflect the SDP of a doctor.

4.2 Doctor-Patient Tie Mining

We will take the preprocessing problem into consideration if a medical dataset for evaluations is too large. This procedure is mainly to filter out those connections

Table 2. Illustration of One Patient's Personal and Diagnostic Information

Patient No.	A73_519	Full Name	User73	Nationality	Han
Age	22	Birth Place	Hunan	CHS	H1,H2
Gender	Female	Position	Student	Affiliation	LA73
Height	164 cm	Weight	49 kg	HBP	100 kPa
LBT	35.5°C	RBT	35.5°C	LBP	65 kPa
Sleep Quality	Normal	Appetite	Normal	Tongue Condition	TC1,TC2
Right Chi	Deep	Left Guan	Float	Urine Condition	Normal
Left Chi	Deep	Right Guan	Deep	Stool Condition	1/(3-4)
Disease History	Null	FDH	Null	Pulse Condition	SS
Doctor No.	1001	NSDS	219	Symptoms	S1, S2
EoD	3	AD-CDs	Level 5	Hospital Location	Fuxingmen Avenue

Table 3. An Instance of Questionnaire Information

DN	AD	MTL	AP	PR	DE	CE	NSCI	EI
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1002	9	3	Good	Yes	Low	Medium	169	5
1003	10	3	Good	Yes	Low	Medium	199	5
1009	8	3	Good	Yes	High	Medium	142	5
1001	7	5	Bad	Yes	High	Poor	108	3
1008	8	4	Medium	No	High	Poor	246	5
1007	7	5	Medium	No	High	Good	113	Null
1005	10	4	Medium	Yes	Medium	Good	166	5
1006	10	5	Medium	No	Medium	Good	209	4
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Note: DN means doctor No., AD means authority of doctors, MTL means medical technical level, AP means the attitude toward patients, PR means price rationality, DE means diagnosis efficiency, CE means curative effect, NSCI means the number of cured cases, and EI means evaluation of intimacy.

that do not stand for doctor-patient cooperation relationships, and accordingly reduce the time and space cost of the TPFG-based method. Constraints of this procedure are listed below.

1) During the period of treatment/cooperation between a_i and a_j, there exists $IR_{ij}^t < 0$ in the time series $\{IR_{ij}^t\}_t$ of IR value.

2) During the period of treatment/cooperation between a_i and a_j, the length of $\{kulc_{ij}^t\}_t$ series does not change. Here, $\{kulc_{ij}^t\}_t$ is used to measure the cooperation degree between a_i and a_j, and $0 \leqslant kulc \leqslant 1$.

3) The duration of the treatment/cooperation between a_i and a_j lasts more than ten days. As we know, a period of treatment is typically within 10 days.

4) A patient knows the diagnosis results at least one day later than his/her doctor.

In this experiment, we use a TPFG-based method to mine doctor-patient relationships. If a pair of patient and doctor complies with the above constraints, we would create one edge from a_i to a_j in the doctor-patient cooperation sub-graph H', and then compute the starting time and the end time between a_i and a_j. After building H', we calculate the probability of every edge in H' using TPFG model. We select symbol θ as the threshold telling whether one doctor-patient relationship is true or not. The greater the θ value is, the higher the mining accuracy is, and the lower the recall rate is. In the experiment, we select $\theta = 0.8$ and extract totally 1 180 doctor-patient relationships. The mining accuracy of doctor-patient relationships extracted by our TPFG-based method is 72.4%.

4.3 Effectiveness of RSVM Algorithm

In the training process for Ranking SVM (RSVM), we employ an open-source SVM tool, SVMlight[26]. The model regards users' answers as a training dataset, and then generates training data by means of a feature extraction procedure. After training the model, we obtain the ranking function $f = (\boldsymbol{\omega}^*, \boldsymbol{y}')$ where every feature has its own learned weight value. An example is illustrated in Table 4.

Table 4. Feature Values Given by RSVM

No.	Feature	Value
1	*M-index*	5.012 5
2	*DomainRel*	2.638 2
3	*Activity*	1.802 4
4	*Uptrend*	−0.674 5

Table 4 shows that the feature *M-index* has the greatest effect on our ranking model for computing the authority degree of doctors, which indicates a good doctor must be a medical expert. $DomainRel = 0$ indicates the matching degree between a patient's disease type and a doctor's diagnosis scope has no effect on the doctor's AD-CDs ranking score. When $DomainRel > 0$, the bigger the *DomainRel* value is, the higher the doctor's AD-CDs ranking score. Here, $DomainRel = 2.638\,2$ means that patients tend to give high scores to those doctors whose diagnosis scopes are consistent with their disease types. $Activity = 1.802\,4$ (> 1) indicates that patients prefer to visit doctors who have more medical activities than others. Among these feature values, the *Uptrend* value may be negative (e.g., $-0.674\,5$), which would mean that a doctor has fewer medical achievements than before.

To improve ranking speed, the model computes all features except *DomainRel* off-line for every doctor, and stores them in a relational database. The system will fail to accomplish authority-degree sorting of all doctors if the database is too large. Hence, as a new strategy, we first divide the whole database into many sub-datasets, and then resort them using our ranking model on these sub-datasets. Compared with sorting the entire dataset, this strategy not only improves the on-line computation efficiency, but also obtains the sorting results with a smaller error rate.

4.4 Recommendation Performance

In this evaluation experiment, we adopted four indexes, $P@5$, $P@10$, $P@15$, and MAP, to evaluate the ranking algorithm for the authority of doctors. $P@k$ indicates precision rates of the first k results that the system outputs towards inputted query keywords, and is defined in (11).

$$P@k = \frac{\text{number of disease cases in the first } k \text{ results}}{k}. \tag{11}$$

MAP denotes an average precision (AP) corresponding to query keywords of every disease. Specifically, given a query keyword, the average precision value will be computed according to the precision of the first k results. Namely, MAP is the average value of AP in the whole testing dataset, where AP is described in (12).

$$AP = \frac{\sum_{k \text{ is relevant}} P@k}{\text{number of relevant disease cases}}. \tag{12}$$

Our paper utilizes patients' questionnaire information as a test dataset, and compares RWR-Model's recommendation precision with that of the baseline algorithms of the traditional RSVM method and IDR-Model. The comparative evaluation results are shown in Fig.2. From this figure, we see that the recommendation precision of our RWR-based approach is better than that of both the RSVM method and IDR-Model[10]. Both MAE and RMSE metrics are used to measure the recommendation quality. From Table 5, the proposed RWR-based approach outperforms the other two methods.

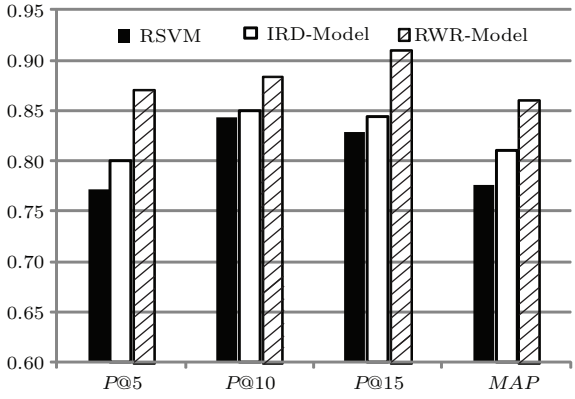

Fig.2. Comparative evaluation results.

Table 5. MAE and RMSE Comparison on Real-World Medical Dataset

Method	MAE	RMSE
RSVM	0.285 34	0.328 65
IDR-Model	0.274 25±(0.002 31)	0.312 64±(0.003 00)
RWR-Model	**0.226 56±(0.004 73)**	**0.211 74±(0.003 29)**

5 Conclusions

In this paper, we tried to systematically investigate the problem of doctor recommendation in medical social networks, and proposed a novel hybrid and multi-layer architecture (namely iBole) to solve it. In iBole, we first mined doctor-patient relationships in a real medical network via a TPFG model. Next, we extracted four features from the network and designed an algorithm based on RSVM for ranking nodes. Finally, we presented a doctor recommendation model via the random walk with restart method. Compared with the baseline methods, RSVM and IDR-Model, our proposed RWR-Model has better recommendation precision. Experimental results show that our proposed recommendation method can help patients find the most appropriate doctor to diagnose their diseases, and that it is a practical technology for intelligent medical information service.

References

[1] Baeza-Yates R, Ribeiro-Neto B. Modern Information Retrieval. Addison Wesley, 1999, pp.98-105.

[2] Salton G, Wong A, Yang C S. A vector pace model for automatic indexing. *Communications of the ACM*, 1975, 18(11): 613-620.

[3] Wang C, Han J W, Jia Y T, Tang J, Zhang D, Yu Y T, Guo J Y. Mining advisor-advisee relationships from research publication networks. In *Proc. the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, July 2010, pp.203-212.

[4] Yang Z, Tang J, Zhang J, Li J Z, Gao B. Topic-level random walk through probabilistic model. In *Lecture Notes in Computer Science 5446*, Li Q, Feng L, Pei J, Wang S X, Zhou X F, Zhu Q M (eds.), Springer Berlin Heidelberg, 2009, pp.162-173.

[5] Macdonald C, Ounis I. Voting for candidates: Adapting data fusion techniques for an expert search task. In *Proc. the 15th ACM International Conference on Information and Knowledge Management*, November 2006, pp.387-396.

[6] Gong J B, Tang J, Fong A C M. ACTPred: Activity prediction in mobile social networks. *Tsinghua Science and Technology*, 2014, 19(3): 265-274.

[7] Hu L, Song G H, Xie Z Z, Zhao K. Personalized recommendation algorithm based on preference features. *Tsinghua Science and Technology*, 2014, 19(3): 293-299.

[8] Shen Y L, Jin R M. Learning personal + social latent factor model for social recommendation. In *Proc. the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, August 2012, pp.1303-1311.

[9] Salakhutdinov R, Mnih A, Hinton G. Restricted Boltzmann machines for collaborative filtering. In *Proc. the 24th International Conference on Machine Learning*, June 2007, pp.791-798.

[10] Gong J B, Sun S T. Individual doctor recommendation model on medical social network. In *Proc. the 7th ADMA*, Part II, December 2011, pp.69-81.

[11] Yang Z, Tang J, Wang B, Guo J Y, Li J Z, Chen S C. Expert2Bólè: From expert finding to Bólè search. In *Proc. the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, June 28-July 1, 2009, pp.1-4.

[12] Tang J, Sun J M, Wang C, Yang Z. Social influence analysis in large-scale networks. In *Proc. the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, June 28-July 1, 2009, pp.807-816.

[13] Kuhn H W. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 1955, 2(1/2): 83-97.

[14] Karimzadehgan M, Zhai C X, Belford G. Multi-aspect expertise matching for review assignment. In *Proc. the 17th ACM Conference on Information and Knowledge Management*, October 2008, pp.1113-1122.

[15] Mimno D, McCallum A. Expertise modeling for matching papers with reviewers. In *Proc. the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, August 2007, pp.500-509.

[16] Karimzadehgan M, Zhai C X. Constrained multi-aspect expertise matching for committee review assignment. In *Proc. the 18th ACM Conference on Information and Knowledge Management*, November 2009, pp.1697-1700.

[17] Hartvigsen D, Wei J C, Czuchlewski R. The conference paper-reviewer assignment problem. *Decision Sciences*, 1999, 30(3): 865-876.

[18] Tang J, Wu S, Sun J M, Su H. Cross-domain collaboration recommendation. In *Proc. the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, August 2012, pp.1285-1293.

[19] Küçüktunç O, Saule E, Kaya K, Çatalyürek Ü V. Diversifying citation recommendations. *ACM Transactions on Intelligent Systems and Technology*, 2015, 5(4): 55:1-55:21

[20] Tang J, Jin R, Zhang J. A topic modeling approach and its integration into the random walk framework for academic search. In *Proc. the 8th ICDM*, December 2008, pp.1055-1060.

[21] Feng W, Wang J Y. Incorporating heterogeneous information for personalized tag recommendation in social tagging systems. In *Proc. the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, August 2012, pp.1276-1284.

[22] Kschischang F R, Frey B J, Loeliger H A. Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, 2001, 47(2): 498-519.

[23] Tang J, Zhang J, Yao L M, Li J Z, Zhang L, Su Z. ArnetMiner: Extraction and mining of academic social networks. In *Proc. the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, August 2008, pp.990-998.

[24] Tang J, Fong A C M, Wang B, Zhang J. A unified probabilistic framework for name disambiguation in digital library. *IEEE Transactions on Knowledge and Data Engineering*, 2012, 24(6): 975-987.

[25] Gong J B, Lu S L, Wang R, Cui L. PDhms: Pulse diagnosis via wearable healthcare sensor network. In *Proc. the 2011 IEEE International Conference on Communications*, June 2011.

[26] Joachims T. Training linear SVMs in linear time. In *Proc. the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, August 2006, pp.217-226.

Ji-Bing Gong is an associate professor at the School of Information Science and Engineering, Yanshan University, Qinhuangdao. He received his Ph.D. degree in computer architecture from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, in 2012. He is a member of CCF. His research interests include data mining, social networks, and machine learning.

Li-Li Wang is a master student at the School of Information Science and Engineering, Yanshan University, Qinhuangdao. She is a student member of CCF. Her research interests include data mining, social networks, and machine learning.

Sheng-Tao Sun is an associate professor at the School of Information Science and Engineering, Yanshan University, Qinhuangdao. He received his Ph.D. degree in geo-information systems from the Center of Earth Observation and Digital Earth (CEODE), Chinese Academy of Sciences, Beijing, in 2012. He is a member of CCF. His research interests include expert systems, intelligent information retrieval, semantic ontology, and grid/cloud computing.

Si-Wei Peng is a lecturer at the School of Information Science and Engineering, Yanshan University, Qinhuangdao. He received his Ph.D. degree in computer application technology from Yanshan University in 2013. His research interests include wireless sensor networks and software engineering.