本文引用格式：李颖,王月,郝建军,等.基于互信息的中医症状推荐系统[J].自动化与信息工程,2023,44(5):52-57.
LI Ying, WANG Yue, HAO Jianjun, et al. Chinese medicine symptoms recommendation system based on mutual information[J]. Automation & Information Engineering, 2023,44(5):52-57.

# 基于互信息的中医症状推荐系统*

李颖[1]　王月[2]　郝建军[3]　王嘉锋[3]
（1.东莞中科云计算研究院，广东 东莞 523000
2.广东电子工业研究院有限公司，广东 东莞 523000
3.广州市黄埔区中医院，广东 广州 510700）

**摘要：**针对中医诊断过于依赖医生经验的问题，提出一种基于互信息的中医症状推荐系统。首先，对原始病例数据进行中医症状规范化，构建症状术语字典，使系统输入规范的症状；然后，通过互信息计算症状之间的关联性；最后，利用归一化折损累计增益（NDCG）指标验证症状的推荐效果，获得症状的推荐列表。实验结果表明，该系统能根据一个症状或多个症状获得其他相关联的症状，实现中医症状推荐功能。

**关键词：**中医；互信息；症状推荐；数据挖掘；症状术语字典；归一化折损累计增益
**中图分类号：**TP311　　**文献标志码：**A　　**文章编号：**1674-2605(2023)05-0008-06
**DOI：**10.3969/j.issn.1674-2605.2023.05.008

# Chinese Medicine Symptoms Recommendation System Based on Mutual Information

LI Ying[1]　WANG Yue[2]　HAO Jianjun[3]　WANG Jiafeng[3]
(1.Dongguan Zhongke Institute of Cloud Computing, Dongguan 523000, China
2.Guangdong Electronics Industry Research Institute Co., Ltd., Dongguan 523000, China
3.Guangzhou Huangpu Traditional Chinese Medicine Hospital, Guangzhou 510700, China)

**Abstract:** A Chinese medicine symptom recommendation system based on mutual information is proposed to address the issue of excessive reliance on doctor experience in Chinese medicine diagnosis. Firstly, standardize Chinese medicine symptoms on the original case data, construct a symptom terminology dictionary, and enable the system to input standardized symptoms; Then, calculate the correlation between symptoms through mutual information; Finally, use the NDCG indicator to verify the recommendation effect of symptoms and obtain a recommended list of symptoms. The experimental results show that the system can obtain other related symptoms based on one or more symptoms, and achieve the recommendation function of Chinese medicine symptoms.

**Keywords:** Chinese medicine; mutual information; symptoms recommendation; data mining; dictionary of symptom terms; normalized discounted cumulative gain

## 0　引言

我国中医学博大精深、历史悠久，是现代医疗体系重要的组成部分。中医的诊疗过程包括四诊识别和辨证论治 2 个阶段，即医生先通过望、闻、问、切，结合诊疗经验辨别患者的身体状况和疾病信息；再总结提取相应的症状，得到证候信息，从而做出诊断并给出治疗方案。在症状提取过程中，医生通常根据患者当前症状询问相关联的症状，这个过程非常依赖医生的个人经验，经验较少的医生难以获取准确症状。近年来，随着互联网、人工智能技术的快速发展，相关技术已经应用于中医领域[1-3]，推动了中医现代化发展的进程[4-5]。结合人工智能与大数据技术进行中医辅


* 基金项目：2021 东莞市标准化研究资助项目

AUTOMATION & INFORMATION ENGINEERING

助诊断，推荐与患者当前症状相关联的其他症状，具有十分重要的现实意义。

宋海贝等[6]基于层次聚类和卷积神经网络开发了中医舌像面像辅助诊疗系统，可对舌像和面像进行自动诊断和分析，并将结果实时反馈给用户，达到健康管理的目的。余江维等[7]利用文本挖掘与自动分类技术，通过 TF-IDF 算法进行中医证候的自动分类与量化研究，得到不同证型的证候分布，验证了 TF-IDF 相对熵量化中医证候的可行性。任晋宇等[8]利用数据挖掘和度量学习技术挖掘、整理中医诊疗经验知识，建立病案相似度的计算方法，设计并实现了中医辅助诊疗推荐系统。

推荐系统是互联网领域有效的信息过滤方法，可避免信息过载，实现个性化服务。主流的推荐系统一般采用基于信息内容、基于协同过滤、基于知识、混合的推荐方法[9-10]。近年来，已有许多学者将推荐系统的思想应用于中医症状推荐领域。吴信朝等[11]利用症状之间的余弦相似度确定患者的推荐症状，实现中医症状的推荐功能，解决了人工经验强耦合的问题，能够从较多的相似症状中，筛选并确定患者的推荐症状。曹静[12]提出基于症状关联网络的中医辅助问诊提示症状推荐算法，通过分析中医问诊数据得到下一步问诊提示，提高医生辨证的准确性。

本文利用互信息技术，分析中医症状的相关性，实现根据患者当前症状推荐相关联症状的功能，可辅助医生诊疗，提高医生的工作效率。

## 1　中医症状推荐系统

基于互信息的中医症状推荐系统主要包括在线症状输入模块、症状提取模块、症状关联度计算模块、症状推荐模块 4 部分，系统框图如图 1 所示。

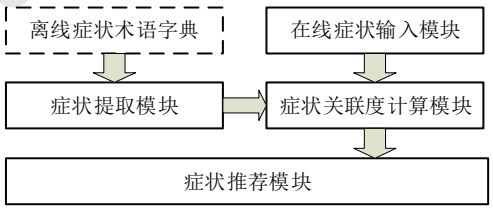


图 1　基于互信息的中医症状推荐系统框图

在线症状输入模块输入患者的当前症状；症状提取模块构建症状病例数据集；症状关联度计算模块计算与输入症状相关联的前 $k$ 个症状；症状推荐模块输出推荐的症状。

### 1.1　症状提取

原始病例数据是描述患者信息的自然语言文本，而基于互信息的中医症状推荐系统的数据处理需使用具体的症状向量，因此需要对病例数据进行症状提取操作。

首先，对原始病例数据进行数据清洗，选取主述及现病史部分文本，通过正则表达式等操作过滤噪声内容，筛选出包含患者症状相关信息的自然语言文本句子。

然后，制定症状术语字典。由于中医症状描述方式众多且尚未有统一的术语字典，同一个症状有多种不同的描述方式。为便于后续处理，将中医症状描述规范化，制定症状术语字典。如鼻腔分泌物清稀、有鼻水这两种症状描述可以规范化为鼻流清涕。先将描述症状的自然语言文本数据转化为结构化数据，并使数据标注尽可能去模糊化；再结合临床病例数据与《中医诊断学》[13]、《中医症状鉴别诊断学》[14]、《常见症状鉴别诊断学》[15]中的症状术语及解释，得到症状术语字典。

最后，利用症状术语字典对病例数据进行症状提取，获得患者的证候信息，构建症状病例数据集。病例数据中的症状提取方法为：1） 通过规范化的症状名及别称进行字符串匹配，匹配相似度利用莱文斯坦距离（一个字符串转成另一个字符串所需的最少编辑操作次数）来衡量；2） 计算 2 个字符串的相似度时，将较长的字符串裁剪成与较短字符串相同长度的多个子字符串，计算各个子字符串与较短字符串的莱文斯坦距离，并以其最小值作为 2 个字符串的相似度。2 个字符串相似度的定义为

$$\mathrm{Sim}_{(a,b)}=1-\frac{L\mathrm{ev}_{(a,b)}}{\min(|a|,|b|)} \tag{1}$$

AUTOMATION & INFORMATION ENGINEERING

$$Lev_{(a,b)}(i,j)=\begin{cases}\max(i,j), & \min(i,j)=0\\ \min\begin{cases}Lev_{(a,b)}(i-1,j)+1\\ Lev_{(a,b)}(i,j-1)+1\\ Lev_{(a,b)}(i-1,j-1)+l_{(a_i\neq b_j)}\end{cases} & ,\text{其他}\end{cases} \tag{2}$$

式中：$Sim_{(a,b)}$ 为 2 个字符串的相似度，$a$ 为包含症状信息的病例数据自然语言文本句子，$b$ 为症状术语关键字，$Lev_{(a,b)}$ 为 2 个字符串的莱文斯坦距离，$\min(|a|,|b|)$ 为 2 个字符串的最短长度，$i$ 为 $a$ 的第 $i$ 个字符，$j$ 为 $b$ 的第 $j$ 个字符，$l_{(a_i\neq b_j)}$ 为 $a_i\neq b_j$ 时，值为 1，否则为 0。

若病例数据的自然语言文本与某个症状关键字的相似度大于设定阈值，则认为该文本中有这个症状。通过不断迭代学习可得到最优阈值，从而准确提取文本对应的症状。

### 1.2 症状关联度计算

基于互信息的中医症状推荐系统的关键步骤为症状关联度计算，通过症状关联度可获得与当前输入症状相关的其他症状。本文利用互信息算法来计算症状病例数据集的症状关联度，流程如图 2 所示。

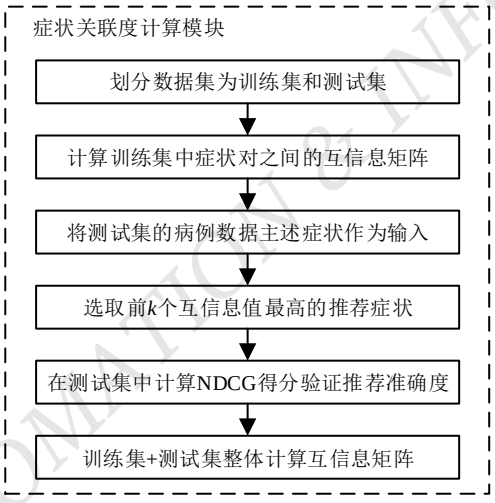


图 2 症状关联度计算流程图

互信息表示变量 $X$ 和 $Y$ 的关联程度，关联程度越高，互信息值越大，计算公式为

$$I(X,Y)=\sum_{x<X}\sum_{y<Y}p(x,y)\log_2\frac{p(x,y)}{p(x)p(y)} \tag{3}$$

式中：$X$ 为某一个随机症状，$Y$ 为另一个随机症状，$p(x,y)$ 为 2 个症状共同出现的概率，$p(x)$、$p(y)$ 为 1 个症状单独出现的概率。

当 $X$ 和 $Y$ 相互独立，没有关联时，$p(x,y)=p(x)p(y)$，$\log_2\frac{p(x,y)}{p(x)p(y)}=0$，互信息值 $I(X,Y)=0$；当 $X$ 是 $Y$ 的一个确定函数，且 $Y$ 也是 $X$ 的一个确定函数，那么传递的所有信息被 $X$ 和 $Y$ 共享，即确定 $X$ 决定 $Y$ 的值，反之亦然。因此，在此情况下互信息与 $X$ 或 $Y$ 单独包含的不确定度相同，即为 $X$ 或 $Y$ 的熵。

首先，以测试集中病例数据主述症状为输入，现病史的症状为真实症状，利用互信息矩阵计算并选取前 $k$ 个互信息值高的推荐症状；然后，利用测试集中推荐症状的归一化折损累计增益（normalized discounted cumulative gain, NDCG）指标来验证推荐准确度，并根据 NDCG 指标来调整算法参数，反复迭代得到最佳参数；最后，合并训练集和测试集，计算症状的互信息矩阵。

NDCG 用于评估推荐结果的效果，取值范围为 0~1，值越大推荐效果越好，计算公式为

$$NDCG=\frac{DCG}{IDCG} \tag{4}$$

$$DCG=\sum_{i=1}^{k}\frac{rel_i}{\log_2(i+2)} \tag{5}$$

式中：$rel_i$ 为位置 $i$ 的推荐症状的相关性评分，$k$ 为推荐列表的长度，$DCG$ 为折损累计增益，$IDCG$ 为推荐结果按照相关性由大到小排列的前 $k$ 个推荐症状组成的列表。

推荐系统返回一个推荐症状列表，每个推荐症状都有相关性的评分值。

### 1.3 症状推荐

通过症状关联度计算模块得到症状病例数据集的互信息矩阵后，新的在线输入症状可利用该互信息矩阵来计算当前症状与数据集中其他症状的互信息值，最后选取前 $k$ 个互信息值高的症状作为当前症状

AUTOMATION & INFORMATION ENGINEERING

的推荐症状。

## 2 实验

### 2.1 实验数据

本文采用的病例数据来自某医院的临床门诊病例数据共 3 312 条。对病例数据中的症状规范化处理后，症状术语字典包含 844 种规范症状，2 232 种别称。

### 2.2 实验结果与分析

首先，将原始病例数据中的主述文本和现病史文本进行数据预处理，划分为包含症状信息的单个文本句子；然后，对文本句子进行症状提取，本文症状提取采用的字符串相似度阈值为 0.7；最后，将症状提取算法得到的症状与真实存在的症状进行对比计算，在病历主诉文本中的准确率为 88.32%，在现病史文本中的准确率为 83.04%。部分文本症状提取结果如表 1 所示。

**表 1 部分文本症状提取结果**

| 病历文本 | 提取症状 | 病历文本 | 提取症状 |
|---|---|---|---|
| 患者感觉鼻塞 | 鼻塞 | 前两天开始咳嗽 | 咳嗽 |
| 晨起鼻胀感 | 鼻胀 | 咽喉咽口水时疼痛 | 咽痛 |
| 每天晚上睡觉会耳鸣 | 耳鸣 | 吃东西时流鼻涕 | 饮食流涕 |
| 吞咽口水耳朵就痛 | 耳痛 | 平素怕冷 | 畏寒 |
| 近 1 个月感冒后声沙 | 声嘶 | 常常感觉憋气 | 呼吸困难 |

本实验采用的 3 312 条病例数据包含 17 875 个症状，431 种症状类别，症状间可两两组合成 9 729 对症状对，其中出现频次最高的前 10 对症状对如表 2 所示。

**表 2 出现频次最高的前 10 对症状对**

| 症状 1 | 症状 2 | 频次/次 |
|---|---|---|
| 失眠 | 烦躁 | 634 |
| 失眠 | 胸闷 | 452 |
| 失眠 | 心悸 | 426 |
| 失眠 | 食欲不振 | 355 |
| 烦躁 | 胸闷 | 335 |
| 胸闷 | 心悸 | 326 |
| 失眠 | 乏力 | 289 |
| 烦躁 | 心悸 | 279 |
| 失眠 | 大便溏 | 247 |
| 失眠 | 眩晕 | 239 |

利用公式(3)计算症状对的互信息值，取得分高的前 10 对症状对如表 3 所示。

**表 3 互信息值最高的前 10 对症状对**

| 症状 1 | 症状 2 | 互信息值 |
|---|---|---|
| 失眠 | 烦躁 | 0.020 5 |
| 胸闷 | 心悸 | 0.016 4 |
| 烦躁 | 胸闷 | 0.013 8 |
| 失眠 | 心悸 | 0.013 0 |
| 失眠 | 胸闷 | 0.012 7 |
| 烦躁 | 心悸 | 0.011 2 |
| 失眠 | 食欲不振 | 0.009 0 |
| 失眠 | 乏力 | 0.006 7 |
| 咳嗽 | 有痰 | 0.006 5 |
| 胃胀 | 嗳气 | 0.006 1 |

由表 3 可知，出现频次高的症状对的互信息值不一定大，这是因为互信息值的计算不仅取决于症状对共同出现的概率，还与每个症状单独出现的概率成反比。如失眠症状出现次数较多，导致包含该症状的症状对的互信息值变小。

为了减少偶然性，将症状病例数据集按 4∶1 随机划分为训练集和测试集。利用训练集数据计算互信息矩阵，将训练集中的症状两两组合，共得到 8 723 对症状对，互信息值最高的前 10 对症状对及互信息值如表 4 所示。

由表 4 可知，训练集中失眠和烦躁症状对的互信息值最高，表示训练集中失眠和烦躁症状关联性相对较高。

AUTOMATION & INFORMATION ENGINEERING

表 4　训练集数据中互信息值高的前 10 对症状对

| 症状 1 | 症状 2 | 互信息值 |
|---|---|---|
| 失眠 | 烦躁 | 0.020 6 |
| 心悸 | 胸闷 | 0.017 2 |
| 烦躁 | 胸闷 | 0.014 6 |
| 失眠 | 心悸 | 0.013 5 |
| 失眠 | 胸闷 | 0.013 4 |
| 烦躁 | 心悸 | 0.011 5 |
| 失眠 | 食欲不振 | 0.008 8 |
| 咳嗽 | 有痰 | 0.006 8 |
| 胃胀 | 嗳气 | 0.006 5 |
| 失眠 | 乏力 | 0.006 4 |

根据症状对的互信息值，构建症状对互信息矩阵。该矩阵是一个 431×431 的二维数组，每一行每一列为一个症状，数值为症状对的互信息值。利用互信息矩阵计算训练集中的推荐症状，具体操作为：将训练集的主述症状作为输入，计算其对应的推荐症状列表；如果输入多于 2 个症状，则将各症状单独输入后得到的推荐列表对应的症状推荐分数相加。测试集随机抽取 5 个病例数据的症状输入与推荐症状（设置为前 10 个）及病例数据中真实出现的症状结果如表 5 所示。

由表 5 可知，推荐的前 10 个症状基本可以涵盖实际症状，仅有个别特殊关联性较小的症状未被推荐，如咳嗽与大便溏的症状对在数据集中只有 69 对，其互信息值较低，前 10 个推荐症状中未给出大便溏的症状。

为了进一步研究不同推荐症状个数对推荐结果的影响，分别计算测试集中5~50个推荐症状的NDCG值及F1分数，结果如表6所示。

由表6可知：随着推荐症状个数增多，推荐结果的 NDCG 值也不断增大，说明增加推荐症状个数有利于数据集中症状关联性较小的症状推荐，可提高频率较少的特殊关联症状推荐的准确度；当推荐症状个数为20时，F1分数最高，说明推荐症状个数为20时，推荐效果最好。

表 5　测试集数据推荐症状及实际症状结果对比

| 输入症状 | 推荐症状 | 实际症状 |
|---|---|---|
| 咳嗽 | 有痰 咽喉痒 痰黄 咽痛 鼻流浊涕 痰白 干咳 气喘 咳痰 鼻塞 | 痰黄 咽痛 大便溏 咽喉痒 咳嗽 |
| 月经过少 | 经色暗紫 失眠 月经后期 经有血块 痛经 乳腺结节 经色淡红 烦躁 湿疹 经期延长 | 经色暗紫 大便干 经有血块 烦躁 |
| 乏力 眩晕 | 失眠 烦躁 心悸 胸闷 食欲不振 头痛 口干渴 夜间多尿 形体消瘦 紧张 | 口干渴 烦躁 紧张 食欲不振 失眠 |
| 失眠 心悸 | 烦躁 胸闷 乏力 食欲不振 胃胀 眩晕 抑郁 大便溏 焦虑 口干渴 | 食欲不振 乏力 体重减少 |
| 小便频数 小便急 | 尿痛 余沥不尽 失眠 鼻腔出血 夜间多尿 鼻流浊涕 鼻塞 多梦 焦虑 脱发 | 余沥不尽 失眠 |

表 6　不同推荐症状个数的 NDCG 值及 F1 分数

| 推荐个数/个 | NDCG 值 | F1 分数 | 推荐个数/个 | NDCG 值 | F1 分数 |
|---|---|---|---|---|---|
| 5 | 0.348 4 | 0.260 8 | 30 | 0.470 2 | 0.270 3 |
| 10 | 0.382 0 | 0.290 8 | 35 | 0.482 0 | 0.255 1 |
| 15 | 0.418 0 | 0.295 0 | 40 | 0.488 7 | 0.244 8 |
| 20 | 0.438 4 | 0.298 8 | 45 | 0.495 0 | 0.233 9 |
| 25 | 0.456 5 | 0.282 9 | 50 | 0.501 0 | 0.221 9 |

根据上述实验结果，设置推荐症状个数为前20个，在症状病例数据集内计算症状对的互信息值，构建互信息矩阵，推荐相关联症状。实验采用5组输入症状，最终的推荐效果如表7所示。

表 7　本文症状推荐系统的推荐效果示例

| 输入症状 | 推荐症状 |
|---|---|
| 失眠 头痛 | 烦躁 胸闷 心悸 食欲不振 眩晕 乏力 胃胀 大便溏 抑郁 口干渴 焦虑 腰背痛 紧张 泄泻 腹胀 多梦 月经过少 腹痛 口苦 畏寒 |
| 腹痛 | 腹胀 泄泻 大便溏 失眠 食欲不振 胃痛 呕吐 肠鸣 嗳气 腹部不适 恶心 喜怒无常 形体消瘦 少腹痛 目昏 烦躁 口干渴 妊娠腹痛 反酸 乏力 |

AUTOMATION & INFORMATION ENGINEERING

续表

| 输入症状 | 推荐症状 |
| --- | --- |
| 咳嗽 有痰 | 咽喉痒 痰黄 痰粘 痰鸣 鼻流浊涕 痰白 失眠 干咳 咽痛 气喘<br>鼻塞 咳痰 胸闷 嗳气 食欲不振 大便溏 少痰 咽中异物感 气促 咽干 |
| 气喘 胸闷 | 心悸 烦躁 失眠 胃胀 乏力 咳嗽 气促 大便溏 有痰 胸痛<br>口苦 抑郁 食欲不振 大汗淋漓 眩晕 四肢冰冷 形体消瘦 头痛 自汗 便秘 |
| 月经紊乱 | 月经后期 经有血块 经期延长 烦躁 经色暗紫 自汗 蚁走感 月经过少 痛经 肥胖<br>月经不来 胸闷 鼻炎 喷嚏 失眠 瘰疬 肌肤疼痛 鼻流浊涕 月经过多 食欲不振 |

由表 7 可知，采用基于互信息的中医症状推荐系统得到的推荐症状基本为输入症状的相关症状。

## 3 结论

本文主要研究从临床病例数据的症状提取基于互信息技术的症状推荐的全过程。实验结果表明，本文提出的基于互信息的中医症状推荐系统可有效推荐当前症状的相关联症状，在医生问诊过程中给予提示，减轻医生因经验不足而导致的诊断困难。同时该系统应用于医院病历系统，有助于医生快速方便地记录症状信息，提高病历录入的效率。后续研究中可不断补充完善病例数据库资源，进一步提高系统的稳定性和可靠性。

## 参考文献


[1] 李艳,杨国庆,双娇月.人工智能在医疗应用中的新进展[J].中国医药导报,2021,18(13):43-46.

[2] 陈挺木.一种疫情防控用服务机器人系统的设计与验证[J].机电工程技术,2022,51(12):241-243.

[3] 刘辉,牛智有.电子鼻技术及其应用研究进展[J].中国测试,2009,35(3):6-10.

[4] 任相阁,任相颖,李绪辉,等.医疗领域人工智能应用的研究进展[J].世界科学技术-中医药现代化,2022,24(2):762-770.

[5] 文杭,黄丽,刘江,等.人工智能技术在中医临床诊疗中的应用研究进展[J].中国医药导报,2021,18(8):42-45.

[6] 宋海贝,温川飙,程小恩.基于 AI 的中医舌象面象辅助诊疗系统构建[J].时珍国医国药,2020,31(2):502-505.

[7] 余江维,余泉,张太珍,等.基于 TF-IDF 相对熵的中医证候量化研究[J].世界科学技术-中医药现代化,2015,17(10):1986-1991.

[8] 任晋宇,白琳,钟华.中医辅助诊疗推荐系统设计与实现[J].中国中医药图书情报杂志,2021,45(3):1-5.

[9] XU Hailing, WU Xiao, LI Xiaodong, et al. Comparison study of Internet recommendation system[J]. Journal of Software, 2009, 20(2):350-362.

[10] 郑诚,徐启南,章金平.基于互信息的推荐系统方法研究[J].微电子学与计算机,2018,35(12):76-79;84.

[11] 吴信朝,阮晓雯,陈远旭.一种无监督中医症状推荐方法、装置、设备及介质:CN114743670A[P].2022-07-12.

[12] 曹静.基于复杂网络的推荐算法在中医辅助问诊中的应用研究[D].镇江:江苏大学,2018.

[13] 李灿东.中医诊断学[M].北京:中国中医药出版社,2016.

[14] 姚乃礼.中医症状鉴别诊断学[M].北京:人民卫生出版社,2002.

[15] 朱豫川,郑海军,冯卫华.常见症状鉴别诊断学[M].北京:中医古籍出版社,2001.



作者简介：

李颖，女，1986 年生，博士研究生，工程师，主要研究方向：知识图谱和深度学习在中医药大数据的融合应用。E-mail: liying@casc.ac.cn

王月，女，1996 年生，硕士研究生，工程师，主要研究方向：自然语言处理在中医领域的应用研究。E-mail: wangyue_hit0616@163.com

郝建军，男，1955 年生，教授，主任中医师，主要研究方向：临床内科和中西医结合的临床研究。E-mail: 2217064411@qq.com

王嘉锋，男，1979 年生，大学本科，主任中医师，主要研究方向：中医内科、医院管理。E-mail: 670097078@qq.com


AUTOMATION & INFORMATION ENGINEERING