# 基于“种子概念”的中医医案症状信息提取方法

徐 亮[1],陈 阳[2],陈守强[1],左瑶瑶[1],毕思玲[1],袁 锋[3*]

(1.山东中医药大学第二附属医院,山东 济南 250014;2.山东师范大学 信息科学与工程学院,山东 济南 250014;3.山东省高等学校中医药数据云服务重点实验室/山东管理学院,山东 济南 250001)

**摘 要:目的:**基于种子概念提取方法,建立中医医案症状信息提取方法。**方法:**对中医医案进行分词,通过设立种子概念,统计互信息量,选择大于阈值者为复合种子概念;结合萤火虫算法,统计相邻词语间关联性,获取扩展种子概念。**结果:**获取复合种子概念217个,扩展种子概念68个。**结论:**该方法实现了中医医案症状信息的自动化提取,为中医医案的数据挖掘提供了便利。

**关键词:**中医医案;症状信息;信息提取;种子概念;萤火虫算法

**中图分类号:**R249 **文献标识码:**A **文章编号:**1673-2197(2018)09-0112-03

**DOI:**10.11954/ytctyy.201809040

## A Method for Information Extraction of TCM Symptoms Based on Seed Concept

Xu Liang[1], Chen Yang[2], Chen Shouqiang[1], Zuo Yaoyao[1], Bi Siling[1], Yuan Feng[3*]

(1. The Second Hospital Affiliated to Shandong University of Traditional Chinese Medicine, Jinan 250001, China;
2. School of Information Science and Engineering of Shandong Normal University, Jinan 250001, China;
3. Key Laboratory of TCM Data Cloud Service in Universities of Shandong/Shandong Management University, Jinan 250001, China)

**Abstract: Objective:** To establish the information extraction method for TCM symptoms based on seed concept. **Methods:** Firstly, segment words of TCM records, and set the concepts of "seed"; secondly, calculate the mutual information, and select the concepts above threshold as composite seed concept; finally, get the extended concepts by calculating the correlation between adjacent words based on firefly algorithm. **Results:** We got 217 composite seed concepts and 68 extended concepts. **Conclusion:** The method realize the automatic extraction of TCM symptoms information, which is helpful for data mining of TCM records.

**Keywords:** TCM Records; Symptom Information; Information Extraction; Seed Concept; Firefly Algorithm

信息技术高速发展,各领域进入到大数据时代,数据的存储及处理能力越来越重要,而大数据理念也为各领域的发展带来机遇,中医作为传统医学,同样面临信息整合及处理的问题,中医医案承载着历代医家诊疗经验,是中医药信息化的重要研究对象。中医医案记录了整个诊疗过程,包括从四诊信息采集、辨证到处方用药整个过程,包含了历代医家诊疗经验,在中医药传承过程中具有巨大价值[1]。然而,由于中医医案概念术语专业性较强、文言句式及描述模糊,所以症状信息的提取困难,数据批量处理及挖掘较难进行。为此,相关研究往往在信息录入过程中对其进行严格的格式化,虽然实现标准化,但主观性强且无法保留原始医案。本研究旨在保留原始中医医案的基础上,通过信息提取,实现其标准化,为后续的挖掘提供数据支撑。本文将结合萤火虫算法,对“种子概念”[2-6]的本体提取方法进行优化,讨论中医医案症状信息的提取方法。

## 1 方法

本文对四诊概念的提取主要分以下部分进行:中医医案收集,症状词典的建立,分词、概念提取,概念修剪四部分。

### 1.1 中医医案收集

本文前期通过搜集文献及门诊病历,建立了中医电子病历[7],现共有近现代名老中医医案6 587例。

### 1.2 症状词典的建立

选取种子概念构建领域词典,将其加入到分词工具中,

**收稿日期:**2018-03-28

**作者简介:**徐亮(1989—),男,山东中医药大学第二附属医院住院医师,研究方向为中西医结合心血管疾病防治及中医药信息化。

**通讯作者:**袁锋(1972—),女,山东省高等学校中医药数据云服务重点实验室教授,研究方向为中医药信息化。

避免使用分词工具对领域文本集进行分词时将领域概念切分成散串，给概念抽取和概念关系抽取带来困难，从而提高分词效率。

### 1.3 分词

词是具有特定含义的汉语基础单元，分词是本体学习的前提条件。本文采用 ICTCLAS 分词系统[8]，结合预定义的领域词典对领域文本集进行分词和词性标注。部分分词结果如图 1 所示。


**图 1 分词结果**

### 1.4 概念提取

1.4.1 *种子概念的设立* 种子概念就是某个概念词语中的核心概念，即以种子概念为中心，可以生成若干概念。在中医医案中，舌质、舌苔、脉象、疼痛等概念可以代表一类领域概念，故本文选择“痛”“闷”“眠”“憋”“乏”“梦”“肿”“胀”“麻”“热”“汗”“纳”“口”“呕”“咳”“痰”“凉”“冷”“满”“烦”“沉”“舌”“苔”“脉”4 个概念，作为种子概念。

1.4.2 *复合种子概念提取* 复合种子概念就是以种子概念为中心生成的概念。如以“苔”为种子概念，可以生成“苔黄”“苔厚”“苔腻”等概念，就是复合种子概念。

复合种子概念的提取，是计算种子与前后词的统计量，达到设定阈值，则认为是复合种子概念。本文选择互信息[9](Mutual Information)作为计算词语之间结合紧密度的统计量，公式[10]如下：

$$MI(A_i,A_{i+1}) = \log_2 \frac{P(A_i,A_{i+1})}{P(A_i)P(A_{i+1})} \quad (1)$$

其中，$A_i$ 与 $A_{i+1}$ 为临近的两个词，$P(A_i,A_{i+1})$ 为两个词同时出现的概率，$P(A_i)$ 与 $P(A_{i+1})$ 为各自出现的概率，$MI(A_i,A_{i+1})$ 反映 $A_i$ 与 $A_{i+1}$ 紧密度，$MI(A_i,A_{i+1})$ 与二者的紧密度呈正相关，$MI(A_i,A_{i+1})$ 越高，紧密度越大，设定阈值 $\theta$，当 $MI(A_i,A_{i+1}) > \theta$ 时，$A_i$ 与 $A_{i+1}$ 构成有效的复合种子概念。

1.4.3 *扩展种子概念提取* 扩展种子概念的获取方法分 2 种：一种是借助于本体库，通过本体库中各概念与种子的关系获取概念；另一种方法是计算相邻词频率。由于目前没有成行的中医领域本体知识库，故本文采用第 2 种方法，统计文本中相邻词出现频率，由于计算量庞大，故引用萤火虫算法，对文本进行扫描挖掘。相关概念对应如表 1 所示。

**表 1 扩展概念提取与萤火虫发光行为为对应关系**

| 萤火虫发光行为 | 扩展概念提取 |
|---|---|
| 萤火虫个体 | 侯选词 |
| 个体亮度的大小 | 概念的优劣 |
| 最亮的个体 | 概念 |

萤火虫算法的描述如下[11-14]：定义 1 为相对荧光亮度：

$$I = I_0 e^{-\lambda r_{ij}} \quad (2)$$

$I_0$ 为萤火虫的所处位置的萤火亮度，即目标函数值，目标函数值越优自身亮度越高；$\lambda$ 为光强吸收系数，一般设为常数。

定义 2 为萤火虫 $i$ 与 $j$ 之间的空间距离：

$$r_{ij} = || x_i - x_j || = \sqrt{\sum_{k=1}^{D} (x_i^k - x_j^k)^2} \quad (3)$$

定义 3 为吸引度：

$$B = B_0 e^{-\lambda r_{ij}^2} \quad (4)$$

$B_0$ 为最大吸引度，一般设为常数 1；$\lambda$ 及 $r_{ij}$ 意义同上。

定义 4 为萤火虫 $i$ 向萤火虫 $j$ 移动的位置更新由式：

$$x_i = x_i + \beta(x_j - x_i) + \alpha(rand - 1/2) \quad (5)$$

其中 $x_i$、$x_j$ 为第 $i$ 个萤火虫和第 $j$ 个萤火虫所处的空间位置；$\alpha$ 为步长因子，常被定义为常数；$rand$ 是服从均匀分布的随机因子。

萤火虫优化算法流程如下：

Step1：初始化基本参数；

Step2：随机生成个体的初始位置，计算萤火虫的目标函数值，定义为 $I_0$；

Step3：由式(2)、(4) 确定萤火虫的相对亮度 $I$ 和吸引度 $\beta$；

Step4：由式( 5) 确定萤火虫新的位置，对处在最好位置的萤火虫进行随机干扰，对其位置进行微调；

Step5：根据萤火虫新的位置计算荧光亮度；

Step6：当满足搜索条件则转 Step7；否则搜索循环次数加 1，转 Step3，执行下一次搜索；

Step7：输出全局极值点和最优个体值，即扩展种子概念。

### 1.5 概念修剪

挖掘出的概念中，某些互信息较大的词串不构成术语，并不具有中医症状信息，本文采用概念修剪法删除非领域概念，采用中文文本分类语料库 TanCorpV1.0 中的 12 个类别的语料库作为背景语料库，通过对比概念在背景语料库的频率，确定是否为中医领域术语，删除非领域概念。

## 2 结果

通过复合种子概念提取算法提取出 217 个复合概念，其中“胸闷”最多，出现 943 次。见表 2。

**表 2 复合种子概念提取结果(部分)** (n)

| 概念名称 | 出现频次 | 概念名称 | 出现频次 | 概念名称 | 出现频次 |
|---|---|---|---|---|---|
| 胸闷 | 943 | 头痛 | 290 | 咳嗽 | 149 |
| 舌红 | 766 | 口干 | 278 | 浮肿 | 146 |
| 乏力 | 751 | 苔薄黄 | 278 | 憋闷 | 146 |
| 脉沉 | 664 | 闷气 | 248 | 头胀 | 130 |
| 气虚 | 644 | 心烦 | 234 | 舌淡 | 128 |
| 脉弦 | 619 | 沉弦 | 206 | 苔黄腻 | 124 |
| 苔白 | 558 | 脉弱 | 195 | 睡眠 | 119 |
| 眠差 | 461 | 脉弦细 | 179 | 腰痛 | 116 |
| 苔薄白 | 444 | 苔少 | 170 | 舌暗 | 115 |
| 憋气 | 430 | 纳差 | 169 | 麻木 | 107 |
| 疼痛 | 363 | 多梦 | 168 | 背痛 | 100 |
| 脉沉细 | 306 | 隐痛 | 165 | 心绞痛 | 89 |
| 胸痛 | 298 | 脉细 | 162 | 苔黄厚 | 80 |

将阈值设定为 5，得到 68 项扩展概念，概念“冠心病”频

率最高，其频次达795。得到扩展概念如表3所示。

**表3 扩展概念提取结果(部分)** (n)

| 概念名称 | 出现频次 | 概念名称 | 出现频次 | 概念名称 | 出现频次 |
|---|---|---|---|---|---|
| 冠心病 | 795 | 入睡 | 148 | 后背 | 80 |
| 高血压 | 612 | 心肌 | 136 | 易怒 | 77 |
| 不适 | 384 | 脱漏 | 130 | 耳鸣 | 76 |
| 劳累 | 311 | 心肌炎 | 109 | 胃脘 | 73 |
| 阳亢 | 292 | 恶心 | 101 | 心功能 | 63 |
| 血压 | 190 | 胸部 | 93 | 心悸 | 60 |
| 齿痕 | 184 | 感冒 | 90 | | |
| 困难 | 150 | 背部 | 83 | | |

## 3 讨论

本文将种子概念提取方法引用到中医医案的摘要提取中，挖掘出复合种子概念；并结合萤火虫算法对扩展种子的提取过程进行优化，提高运行效率。

扩展种子概念提取还有另一种方法，即依托于现有的本体库，通过本体概念之间的相互关系，挖掘出扩展种子概念。在目前尚未有中医本体库的情况下，由于中医术语的专业性较强，结合萤火虫算法的扩展种子概念提取方法是最佳选择。其中，萤火虫算法(Firefly Algorithm，FA)，最早由剑桥学者Yang[15]于2008年提出，通过模仿萤火虫行为，由此构造出的一种随机寻化算法[16-18]。萤火虫算法的原理是[19-20]：将问题的解模拟成萤火虫个体，在本文中相当于相邻词语，萤火虫有各自的荧光素值与感知半径。其中荧光素值即萤火虫的亮度，用来衡量解的个体位置的优劣，荧光素值越大，则个体吸引力越强，即解更优，在本文中相当于相邻词语的关联性。

由于中医医案术语具有模糊性的特点，在对中医医案进行数据挖掘时，必须进行较复杂的数据预处理，使其转化为可用于统计的数据化信息。多数研究选择通过建立中医症状、诊断数据库的方法对信息录入进行规范化，即用户在录入医案时只能选择库中现有的数据，此方案将医案转换成可以直接用于统计的规范化词语，省去了数据预处理过程，但是此方案存在弊端：①未能将原始的中医医案保存下来；②由于每次研究的针对性不同，规范化的信息录入针对性也不同，致使在不同研究中不得不进行重复的医案录入过程，耗费大量时间与精力。中医概念提取技术的运用，将实现中医医案症状、舌苔、脉象等四诊信息的自动化提取，既能保存原始的中医医案资料，避免医案的重复录入，又能提高数据挖掘的效率。

中医医案的不断积累、更加完善的中医概念字典的建立以及更优化的算法应用可使中医药领域的概念提取技术更加完善，结果更精确，为中医医案的数据挖掘提供更坚实的支持。

**参考文献：**

[1] 张帆，刘晓峰，孙燕. 中医医案文献自动分词研究[J]. 中国中医药信息杂志，2015，22(2)：38-41.

[2] MOLDOVAN D，GIRJU R，RUS V. Domain-specific knowledge acqui-sition from text [C]. Proc. of the Sixth Conference on Applied Natural Language Processing，2000：268-275.

[3] 王红滨，刘大昕，王念滨，等. 基于遗传算法和种子概念的本体概念提取算法[J]. 系统工程与电子技术，2010，32(11)：2465-2469.

[4] 何超，张玉峰. 融合领域本体的中文文本语义特征提取算法研究[J]. 情报理论与实践，2013，36(9)：96-99.

[5] 汪平仄，曹存根，王石，等. 一种迭代式的概念属性名称自动获取方法[J]. 中文信息学报，2014，28(4)：58-67.

[6] 李文杰，穗志方. 基于并列结构的概念实例和属性的同步提取方法[J]. 中文信息学报，2012，26(2)：82-87.

[7] 陈守强，张建民. 中医门诊电子病历的设计与应用研究[J]. 山东中医药大学学报，2002，26(6)：428-429.

[8] 刘群，张华平，俞鸿魁，等. 基于层叠隐马模型的汉语词法分析[J]. 计算机研究与发展，2004，41(8)：1421-1429.

[9] 史益新，邱天爽，韩军，等. 基于混合互信息和改进粒子群优化算法的医学图像配准方法[J]. 中国生物医学工程学报，2015，34(1)：1-7.

[10] 梁健，吴丹. 种子概念方法及其在基于文本的本体学习中的应用[J]. 图书情报工作，2006，50(9)：18-21.

[11] YANG XINSHE. Firefly algorithms for mult-imodal optimization[C]. Proc of the 5th International Symposium on Stochastic Algorithms：Foundations and Applications，2009：169-178.

[12] YANG XINSHE，DEB S. Eagle strategy using lévy walk and firefly algorithms for stochastic optimization [J]. Studies in Computational Intelligence，2010(284)：101-111.

[13] 袁锋，陈守强，刘弘，等. 一种改进的文化萤火虫算法[J]. 计算机仿真，2014，31(6)：261-265.

[14] 刘长平，叶春明. 一种新颖的仿生群智能优化算法：萤火虫算法[J]. 计算机应用研究，2011，28(9)：3295-3297.

[15] YANG XINSHE. Nature-inspired metaheuristic algorithm [M]. Frome：Luniver Press，2008：81-96.

[16] 亢少将. 萤火虫优化算法的研究与改进[D]. 广州：广东工业大学，2013.

[17] 赵鹏军，李会荣，刘晓民. 一种基于全局最优的改进萤火虫算法[J]. 河南科学，2017，35(11)：1729-1734.

[18] 郝晓莹，贺兴时，薛菁菁. 一种粒子群-萤火虫算法的参数优化方法[J]. 西安工程大学学报，2017，31(5)：695-700.

[19] 左仲亮，郭星，李炜. 一种新颖的改进萤火虫算法[J]. 微电子学与计算机，2017，34(9)：15-19.

[20] 隋永波. 萤火虫算法的理论分析及应用研究[D]. 湘潭：湘潭大学，2017.

(编辑：尹晨茹)