中医知识网络可视化的研究与实现

刘清晨 刘鹄 门畅

# 一、背景介绍

中医作为我国的传统医学，有着悠久的历史和独特的体系结构。长期以来，中医奉行的是“天人合一”的思想和系统的整体辩证思想，在当今社会有其独到的价值。在医疗技术飞速发展的当代社会，医学界越来越意识到，对于越来越多的健康问题尤其是慢性疾病，只有从生理、社会、文化、经济、心理等多个方面考虑，从系统的层次去研究，才是攻克这些问题的方向与关键。而中医的思想，恰好与这“生理-心理-社会”不谋而合。

一方面，正如世界卫生组织(WHO)在2014年通过的传统医学决议中所说，“在卫生保健和疾病预防与治疗方面，尤其针对慢性病，传统医学有很长的历史”；中医等传统医学的价值越来越得到重视。另一方面，由于中医长期以来是一种个性化的经验学科，“证”“候”等概念复杂和模糊，大规模的循证医学分级可靠性又较低，因此现代中医的作用又受到争议。如何更好的系统的整理中医的思想，形成可被检验的系统的标准化知识网络，成了当代中医的一个重要的发展方向。

## 中医的系统与网络化研究

20世纪90年代开始，一门着眼于整体，按照中医的全局理念、现代医学的分子技术整合交叉的新的医学理念开始形成，这一理念后来被称为系统医学。系统医学认为，机体内无数的结构是一个复杂系统。这也就为研究中医的理论提供了一个新的切入点。长期以来，中医循证的困难在于无法在微观层面上对于疾病和治疗的机制做出合理可信的解释，而系统医学的思想将机体看作复杂系统，将中医的理论和西医的微观证据结合起来，这也给了研究人员新的启发。

清华大学李梢教授等在2009年(李梢, 2009)提出了新思路，建立病症—组分（因素）—药物之间的网络，从网络角度入手解决中医的循证问题。采用网络药理学、系统生物学的方法，在“寒热”类生物与方剂研究中取得新进展。

中国中医研究院信息所在2002年开始，借鉴一体化医学语言系统(UMLS)，开始构建中医药药学语言系统(TCMLS)(贾李蓉，刘丽江, 2012)。这一系统目前已经初步在文本挖掘，资源检索领域有了应用。TCMLS已经于2014年成为国际标准。贾李蓉和于彤等整理的TCMLS的整体系统(贾李荣，于彤，李海燕，刘静，刘丽红，朱玲，董燕，高博，崔蒙, 2014)网络结构如图1.1。

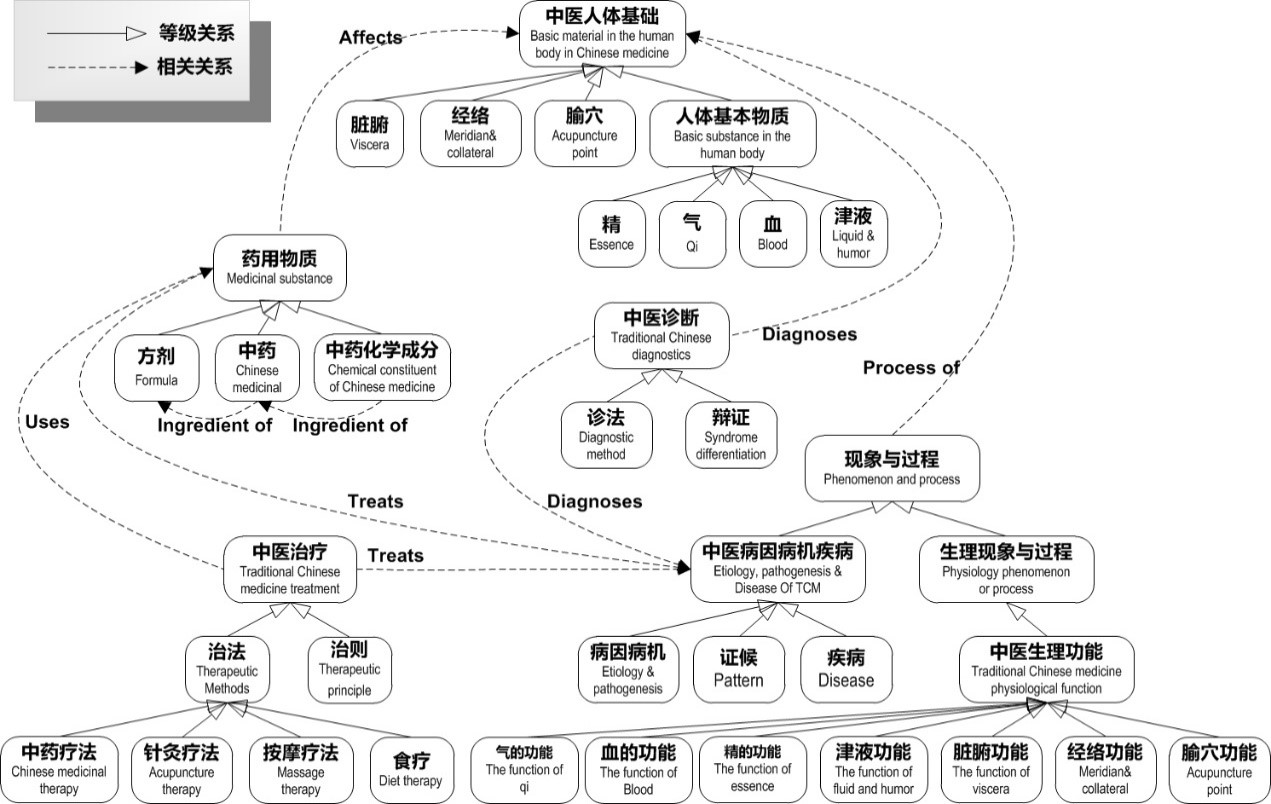


图1 TCMLS整体结构

这些领域的研究，都说明了中医知识的系统化和网络化研究正在不断深入化，但我们也发现这些研究需要大量人力物力时间精力。如何更好的整理和学习研究中医体系，计算机的数据挖掘和机器学习可以提供很好的帮助。

### 1.1中医知识的量化体系

在医学知识体系的构建获取方面，主要源自两个大的方向。文献理论和样本数据。余辉等在《医学知识获取与发现的研究》(余辉, 2003)中，阐释了这两种来源的关联。下面采用关系图谱进行展示中医知识获取的关系，该关系图谱借鉴了上述文献中阐述的理论加以完成。

在图1.2中，医学专家知识获取属于理论出发的演绎，而计算机辅助医学知识发现则是样本出发的归纳。“人”作为个体存储和理解“数据—知识”体系的能力是有限的，因此，随着计算机技术的发展，这种在知识库中学习医学经验或知识的任务，越来越多的落在了计算机的数据挖掘和机器学习上。

计算机对于中医相关数据库的学习和挖掘，主要体现在对于中医的书籍文章和案例等文本理解，现有中医药数据库的存储和整理，以及其他医学数据和理论的学习。

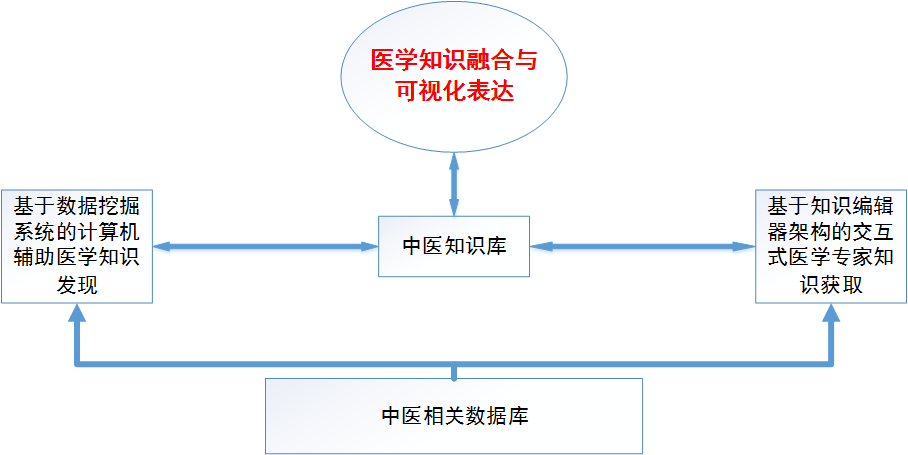


图2 医学知识获取的关系图谱

计算机技术如自然语言处理等的应用，在一定程度上量化了上述所说的中医知识库，更规范化了数据的存储和相互之间的关联。但是，单纯的利用计算机寻找和归纳医学知识，只是知识的获取和发现环节。而对知识表达进行可视化，系统化，模型化的知识表达，可以更好的反馈给医学工作者，从而既能提高计算机学习的准确性，又能给相关行业的研究人员更大的启发。中医知识网络的可视化研究，正是着眼于这一问题。

### 1.2知识网络的可视化表达

知识网络的构建，离不开计算机对于海量的文本信息的处理，但是计算机的处理结果是非直观的。可视化的研究，就是充分利用人体的视觉感知、关联和推理的能力，将计算机学习处理的信息展现给人看，从而获得更好的反馈和纠正。具体的环节和过程可以用框架图3(唐家渝，刘知远，孙茂松, 2013)来展现。

在可视化呈现这部分，又可以细分为基于文本内容的可视化，基于相互关系的可视化以及基于不同类别信息的可视化(唐家渝，刘知远，孙茂松, 2013)，其核心都是将复杂的高维的信息通过一定的方法平面化低维化有利于人视觉的接收与处理。

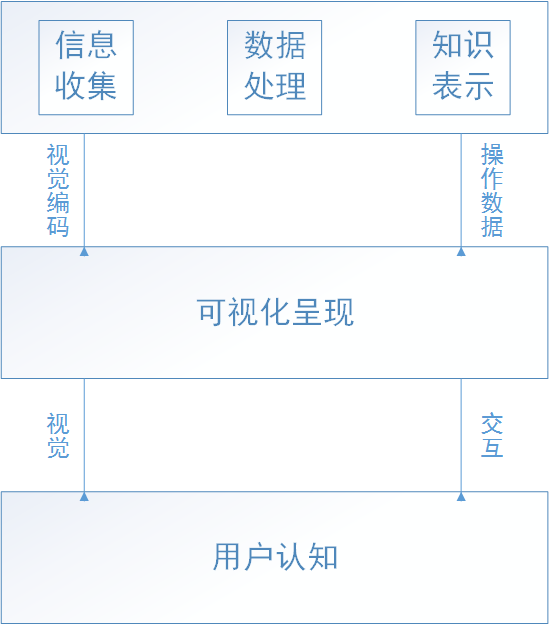


图3 知识网络可视化的交互框图

## 2. 国内外研究概况

### 2.1自然语言处理(NLP)的发展

自然语言处理已经被广泛的运用到医学的各个领域，尤其是文本挖掘和知识网络构建。我们采用的主要是对中医文献资料进行神经网络的训练得到高维的词向量，这里就需要简要介绍一下词向量(word vector)的发展。

词向量最早是 Hinton 在 1986 年的论文《Learning distributed representa-tions of concepts》(Hinton, 1986)中提出的，核心思想是通过深度网络的训练，将每个词汇表示为约100~200维的向量，从而可以看出词汇之间的关系，进一步应用于词性标注、命名体识别、机器翻译等NLP任务中。近两年词向量引出的研究有很多，Tomas Mikolov(Tomas Mikolov, 2013)将RNN（循环神经网络Recurrent neural Network）应用于深度网络，开发了word2vec代码，改进了训练算法速度，之后又提出可以将短语或者文档也通过向量的形式表示的方法，扩展了应用方向，word2vec不仅在训练出的词向量中考虑到了语义关系，也考虑了语法关系。

### 2.2降维技术在文本中的应用

常用的特征降维方法一种划分分类可以分为线性/非线性降维方法；当然也可以按照监督学习的降维和非监督学习的降维来划分，本文第2章中是以线性/非线性降维的结构来详细介绍各个算法的原理的。

最初的线性降维方法主要有主成分分析(PCA)和古典尺度法(MDS)，这两种方法基础而重要，但是直接应用在文本中并不多，总体而言，流形方法为代表的非线性方法更适用于文本的降维分析。

早在2000年当Sam T.Roweis等提出(Sam T. Roweis, 2000)LLE这一非线性降维方法时，就成功的将这一方法应用到文章向量中，成功的实现了相同或相近主题词在二维平面上语义语法聚类的效果。t-SNE算法的作者Laurens van der Maaten等在2012年的文章(Laurens van der Maaten, 2012)中也成功的应用在了文章主题词的聚类中。

越来越多的应用表明降维技术与词向量等自然语言处理技术越来越紧密的结合在了一起，这也是本课题将降维技术应用在中医词向量中的原因之一。

### 2.3文本可视化研究

文本的可视化主要可以按照基于文本内容，基于文本关系和基于多层面信息这三类来划分，但是这种划分也只是很粗略的，实际应用中，经常需要对特定的可视化目标问题设计特定的可视化方法。常用的可视化方法有以下一些。

如考虑文本的出现的频率和词袋模型，最常用的可视化方法就是“标签云”；考虑文本的结构和相互之间的关系，常用树状结构的word tree；如考虑文献的引用关系，还可以采用FP-tree等结构。而我们的知识网络由于词汇量巨大，通常采用聚类文本地图的方法进行展示。

可视化方法发展到今天越发层出不穷，越来越强调交互性和时间空间等特征，一些传统的方法也在不断的融合之中。但是目前来看，文本可视化依然有以下几项有待研究和突破。

(1)缺乏评价标准来衡量可视化方法的准确性，直观性和互动性；

(2)对于大数据信息的处理依然显示的很繁杂；

(3)没有较好的通用方法，需要针对特定数据进行特定的设计。

# 数据准备及初步分析

## 数据获取

构建中医知识网络，首先需要有充分的资料供计算机进行机器学习。在收集数据的过程中，主要采用了Python scrapy爬虫等抓取工具，在Windows环境下，将收集的中医药相关资料主要整理为中医、中药和案例资料这三大类。这三类的收集的资料和搜狗语料库的资料将进行word2vec的中文词向量学习与训练。

主要是在网络上收集了以下类别的资料。

(1) 中医资料：a. 中医经典书籍，有历朝历代医学著作和总结等文献资料；b. 两万多条中医的术语与其详细解释，这部分既可以用作最后的分类中医术语字典材料和依据。

(2) 中药资料：a. 两万条中药词汇术语与详细解释，这部分可以用作分类中药字典的材料；b. 两万六千条中医药方处方的术语和详细成分解释。

(3) 中医案例资料：主要是医生对于药方或者病情的分析案例和医生对于具体病患诊治的过程，共计收集了4000余例。

## 数据处理

### 2.1生成词向量

### 2.2降维

### 2.3计算余弦值

### 2.4过滤

## 初步分析

# 可视化设计及实现

## 整体设计

对于得到的中医知识网络，我们采用力导向布局的节点连接图、弧长链接图、邻接矩阵图进行可视化。其中力导向布局的节点链接图展示完整数据集，提供框选功能，用户可以选出自己感兴趣的数据，在弧长链接图和邻接矩阵图中具体查看；弧长链接图和邻接矩阵图提供按类别排序、协同高亮功能；三个视图都提供悬停高亮、显示数据标签的功能。

## 交互功能

# 四、分析结果

# 五、总结