中医文本数据可视化的研究与实现

刘清晨2015211044刘鹄2015211032门畅2015211061

目录

[一、背景介绍 3](#_Toc454214972)

[1. 中医的系统与网络化研究 3](#_Toc454214973)

[1.1中医知识的量化体系 4](#_Toc454214974)

[1.2知识网络的可视化表达 5](#_Toc454214975)

[2. 国内外研究概况 6](#_Toc454214976)

[2.1自然语言处理(NLP)的发展 6](#_Toc454214977)

[2.2降维技术在文本中的应用 6](#_Toc454214978)

[2.3文本可视化研究 7](#_Toc454214979)

[二、 数据准备及初步分析 7](#_Toc454214980)

[1. 数据获取 7](#_Toc454214981)

[2. 数据处理 8](#_Toc454214982)

[2.1生成词向量 8](#_Toc454214983)

[2.2降维 8](#_Toc454214984)

[2.3计算余弦值 9](#_Toc454214985)

[2.4过滤 9](#_Toc454214986)

[3. 初步分析 9](#_Toc454214987)

[三、 可视化设计及实现 10](#_Toc454214988)

[1. 整体设计 10](#_Toc454214989)

[2. 交互功能 11](#_Toc454214990)

[2.1 力导向图的设计 11](#_Toc454214991)

[2.2 邻接矩阵图的设计 14](#_Toc454214992)

[2.3 弧长链接图的设计 16](#_Toc454214993)

[四、分析结果 18](#_Toc454214994)

[五、总结 18](#_Toc454214995)

# 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。

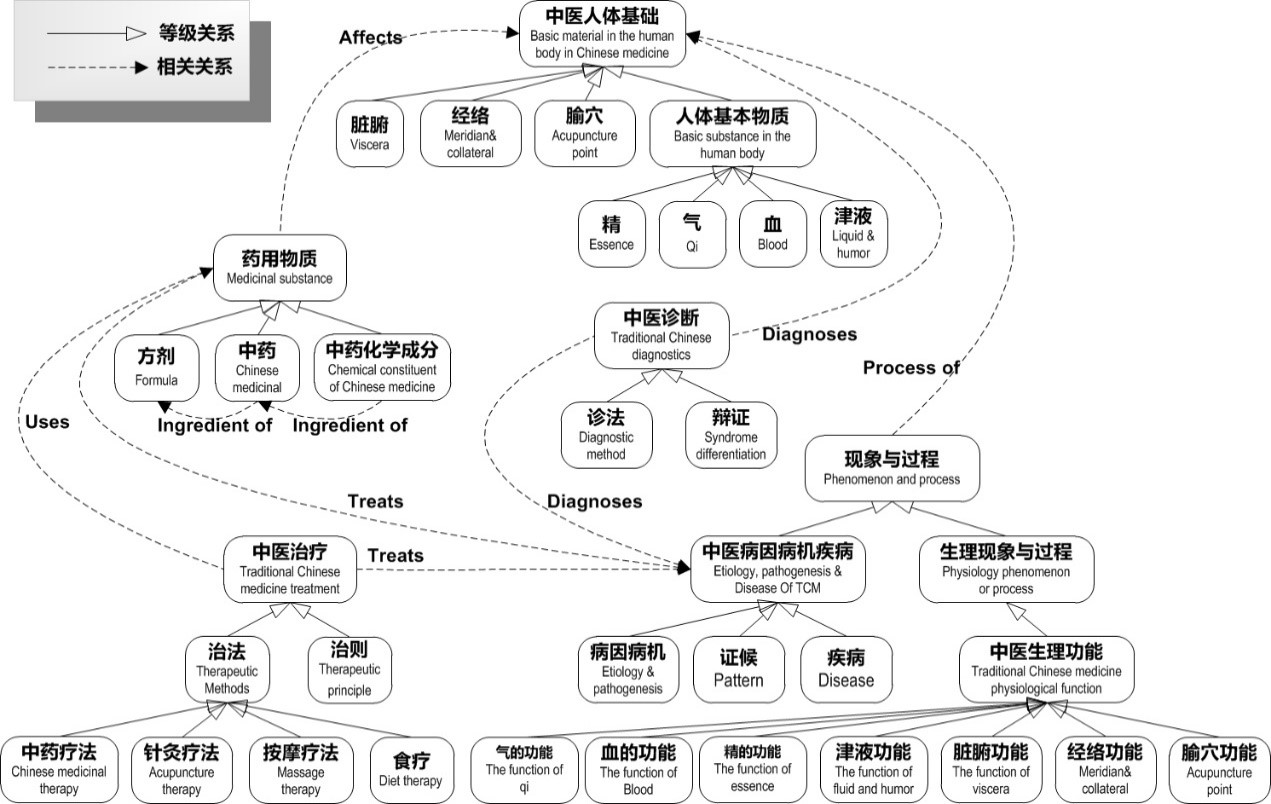


图1 TCMLS整体结构

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。

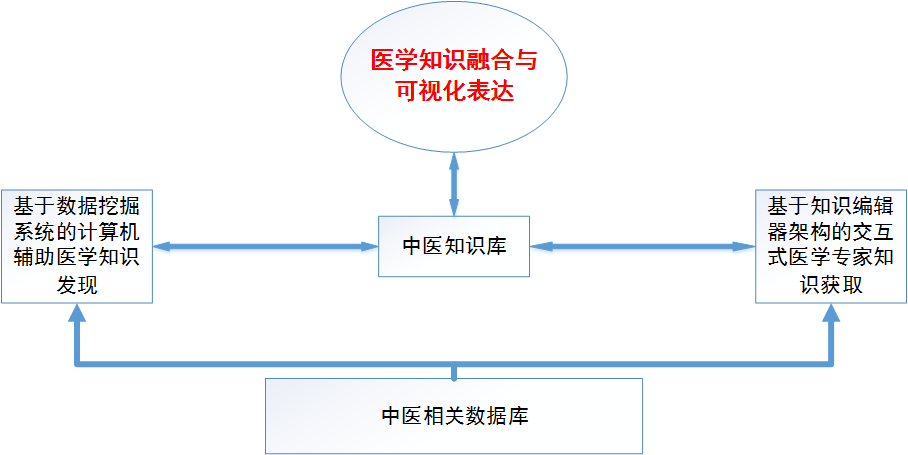


图2 医学知识获取的关系图谱

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。

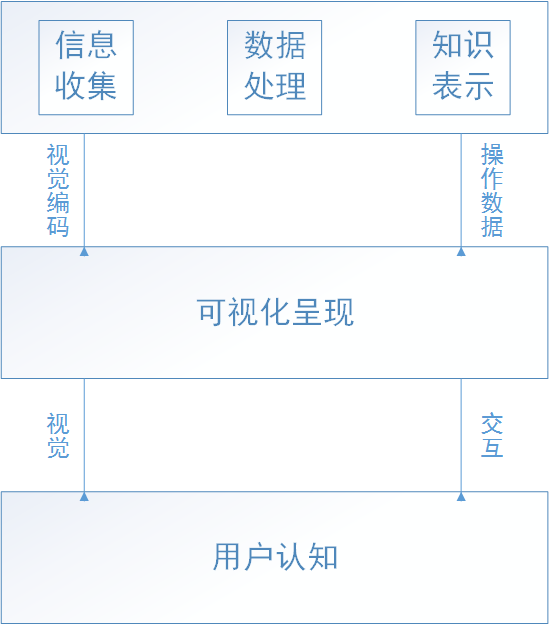


图3 知识网络可视化的交互框图

## 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在jieba分词的基础上进行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。

(4) 搜狗语料库：主要是和医学相关的新闻报道材料，用于训练词向量的主要材料。

## 数据处理

### 2.1生成词向量

使用的是jieba中文分词，然后使用word2vec进行训练，训练后的词向量是100维的表达，初步统计有四千多词。原始数据在/data文件夹的model.txt里面。

对这些词向量的整理时一个复杂的工程，我们小组分工，每个人大约手动标注了1500词，对这些词向量进行分类，主要分为以下几类：

0——不确定属于哪一类

1——和医学无关名词

2——病灶（疾病）

3——表型

4——用药

这部分的工作量比较大，需要人工光一个一个进行标注，我们汇总后将标签统计整理和互相检查校验。标注的原始数据在/annot文件夹中。

### 2.2降维

降维这部分有之前的基础，选择使用t-sne方法，具体来说，使用python的scikit-learn包自带的t-sne方法对高维词向量进行降维，降到二维后方便进行可视化展示。在/script文件夹的generate\_json.py和generate\_label.py脚本负责进行这部分的数据处理，我们在数据处理的过程中将数据尽可能处理成json格式，这样是为了便于可视化展示的时候输入数据。

### 2.3计算余弦值

同样，为了突出表现词向量之间的关系，我们计算了所有向量两两之间余弦的距离。这部分同样使用的是/script的generate\_json.py脚本，不过在使用的时候，2.2的脚本被注释掉了。

生成的是force.json等可以用于展示的json数据。

### 2.4过滤

不是所有的数据都适合展示或者说是有效的网络。我们在最初的尝试中，发现较大的网络展示一是加载比较麻烦，二是有效信息被隐藏在数据点中，起不到展现各个节点关系的作用。于是，我们决定首先筛选有效的词向量，然后再其中以其与其他词向量的余弦值和的大小作为阈值，选择效果比较好的子网络。

在过滤中，我们还注意到以下几个方面：一是我们关注的是不同类别的词汇的关系，因此我们在筛选过程中不考虑同类别的余弦值；二是我们设立的阈值时采用每个词向量和其他类别词向量的余弦值之和，因此需要进行搜索算法来进行寻找子网络。

## 初步分析

在上述的数据清洗和处理中，我们已经可以找到一些很好的相互关联的网络。为了表示的时候有所侧重，我们将数据分为两个json格式来针对不同的视图。Bundle图中不需要展示所有的向量之间的关系，只要关注不同类数据之间的关系就可以了，这是由于bundle图完全展示的时候会比较密集，看不清，因此重点展示不同类数据之间的关系。而力导向图和邻接矩阵相对来说可以展示更复杂的信息，因此使用全部的关系数据。

其中bundle图使用是\_force.json，而力导向图和邻接矩阵使用的是\_force\_all.json.这部分数据放在了tools/d3/experiment文件夹中。

这部分数据一共有20个节点，包括了三类数据，这部分数据使用力导向图展示后基本表达如下图1：

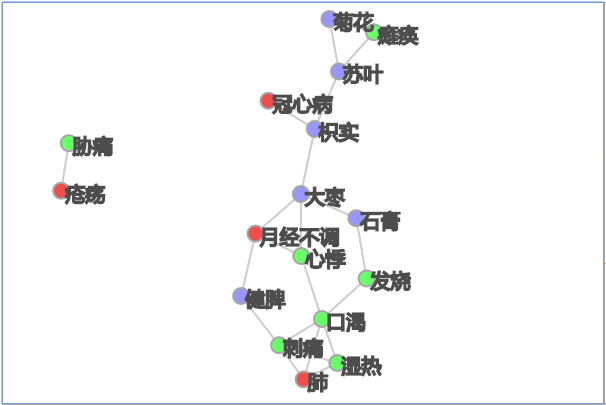


图 1 力导向图表示下的子集数据

# 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。

选择使用的工具是d3.js，为了便于布局和分工，我们小组三人每个人负责一部分视图的表达，其中，刘清晨负责力导向布局的节点连接图，刘鹄负责邻接矩阵图，门畅负责弧长链接图，由刘鹄同学负责汇总。项目整体代码放在了github上面。项目主页是<https://github.com/liuhu-bigeye/med_word_vis>

在可视化的设计中，我们注重细节和交互的表达，对于表达的信息的细节都进行了很多思考和表达，下面的部分将逐步进行介绍。

## 交互功能

### 2.1 力导向图的设计

力导向图的整体设计如下图2

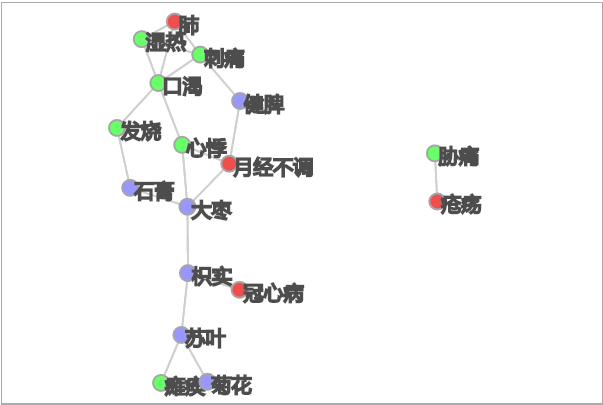


图 2 力导向图整体表达

不同类别的词采用不同颜色的节点来表示，而不同的向量之间的联系采用粗细不同的力线进行连接，线的粗细由向量之间的余弦值的大小决定。

由于力导向图的设计是针对比较复杂的网络，因此在力导向图的设计中，滚动鼠标滚轮可以随时放缩整个网络，便于观察整体网络和局部网络表达。放大后的一个表示如下图3.

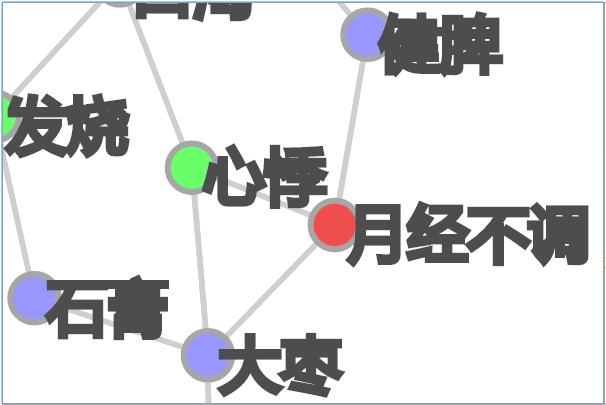


图 3 放大后的网络表达

对于整体网络的表达，默认整体网络是表达在视图的中央，但是在网络的空白处随机用鼠标左键按住移动，整体网络会随之移动，表达效果如下图4

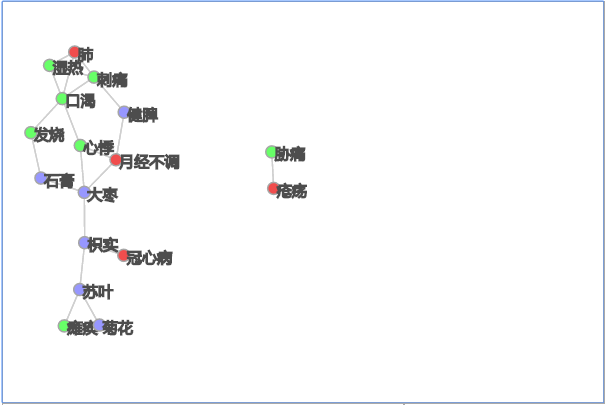


图 4 移动后的网络表达

可以使用鼠标左键点选某个节点，这个节点的文字会变成红色，同时在右侧的data info中会显示该节点的详细信息，如下图5所示。而按下shift键再点解该节点可以取消选中该节点。

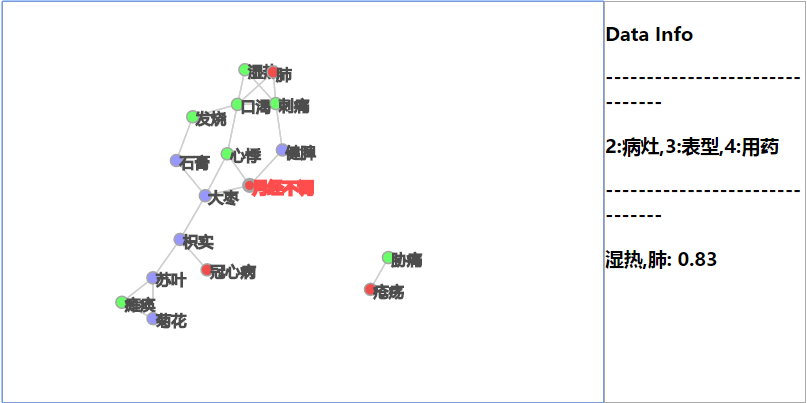


图 5 选中一个节点后显示相关信息

同时按下shift,采用鼠标左键拖动可以选择一个区域里的所有的节点，再按下shift，同样在该区域中选择可以反选，shift的作用就是起到选择或者反选的作用。见下图6

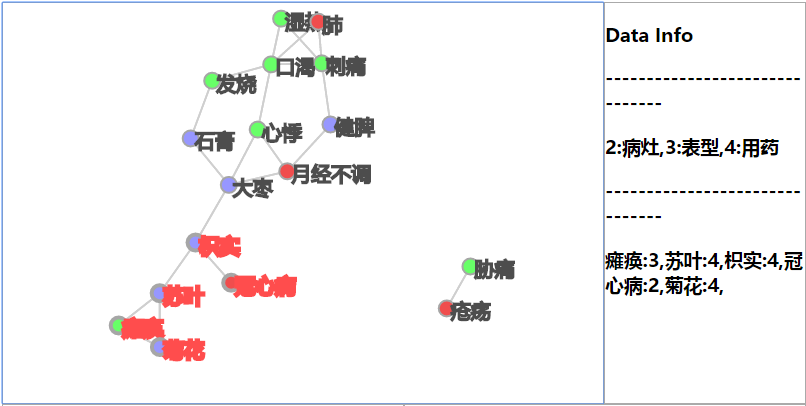


图 6 选中一片区域后的节点信息

### 2.2 邻接矩阵图的设计

邻接矩阵的设计和弧长链接图是联动的，但是邻接矩阵由于便于展现所有节点之间的关系，我们采用不同的颜色色块来表示不同类别之间的关系，这些颜色是利用2.1中不同类别的颜色叠加产生的。

而不同颜色的深浅表示向量之间关系的强弱。如下图7是默认是视图表达。

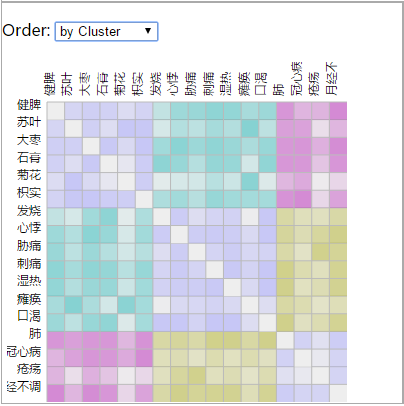


图 7 邻接矩阵图的默认视图

邻接矩阵有三种表达进行切换，图7是按照不同类别进行排序的结果，还可以按照相关性的强弱排序或者按照词汇的拼音进行排序，同时进行了过渡动画的设计。见下图8

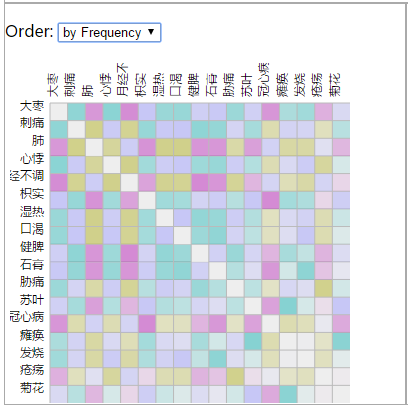


图 8 按照余弦值排序后的邻接矩阵图

鼠标悬停在任意一个方格中，相应的节点会变红，同时在信息栏显示该链接的信息，同时在弧长链接图中会突出显示该链接的弧。见下图9

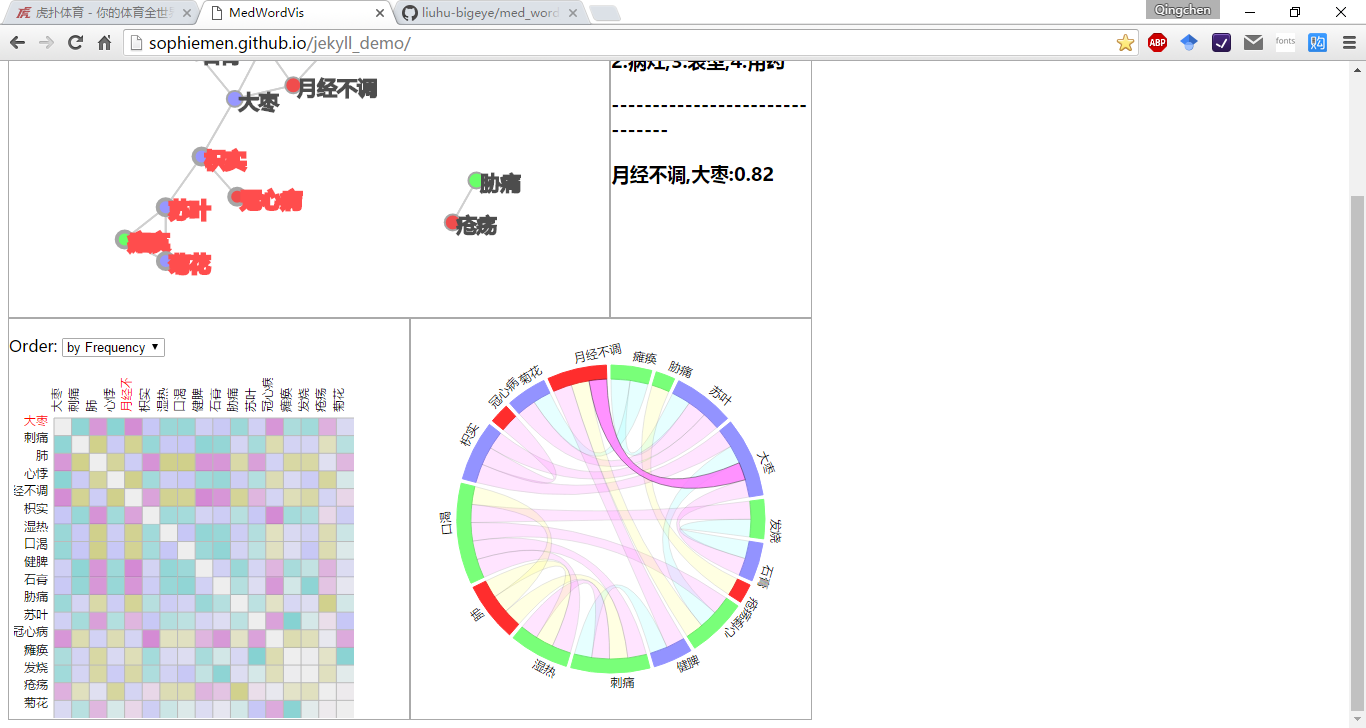


图 9 鼠标悬停后的视图联动和信息表达

### 2.3 弧长链接图的设计

弧长连接图的颜色设计十分出众的。除了不同类别的节点颜色区分，不同的弧的颜色是不同类别的节点颜色叠加形成的。整体的默认表达如下图10

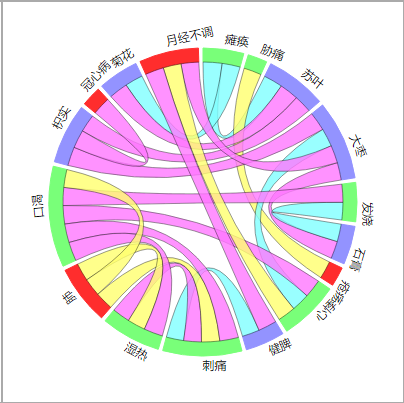


图 10 弧长链接图的默认视图

鼠标悬停在任何弧中都会显示该弧的具体信息，同时也会和邻接矩阵进行联动，见下图11

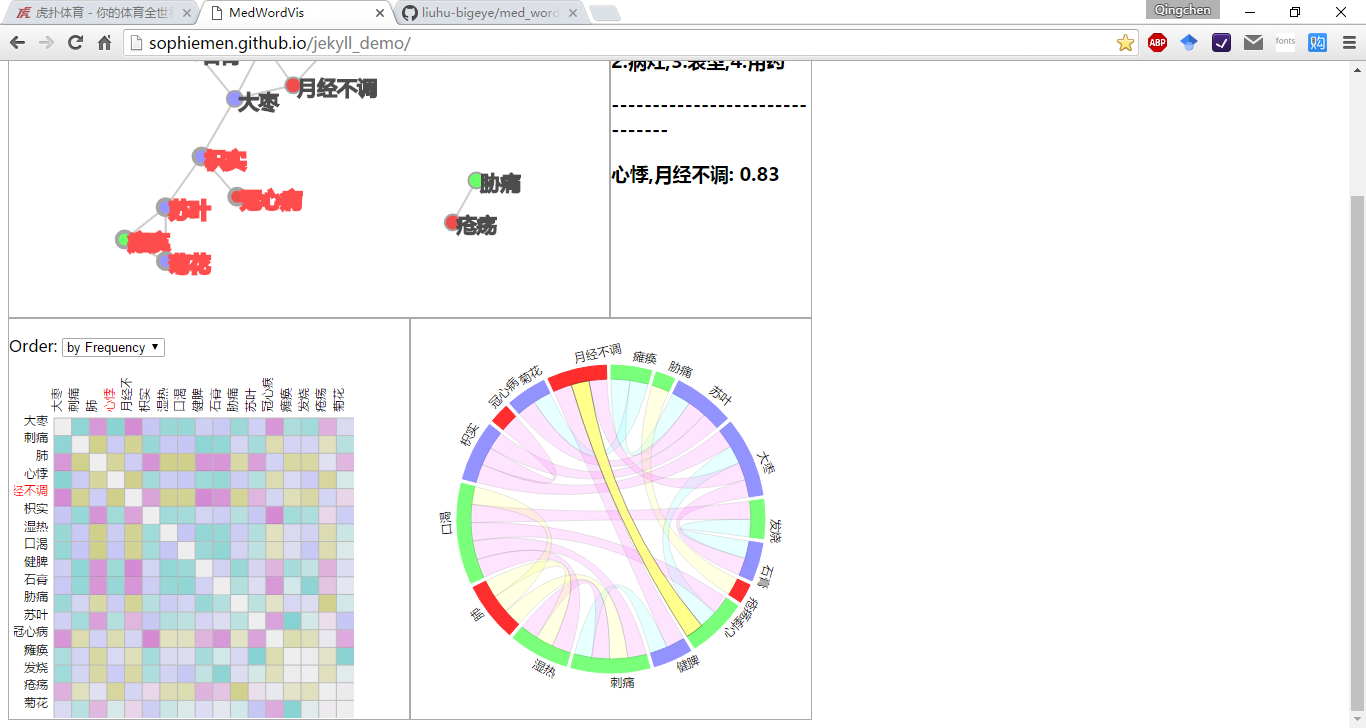


图 11 选中一个弧之后联动视图和信息显示

当然鼠标也可以悬停在节点上，显示和该节点相关的所有弧和所有弧的信息。见下图12

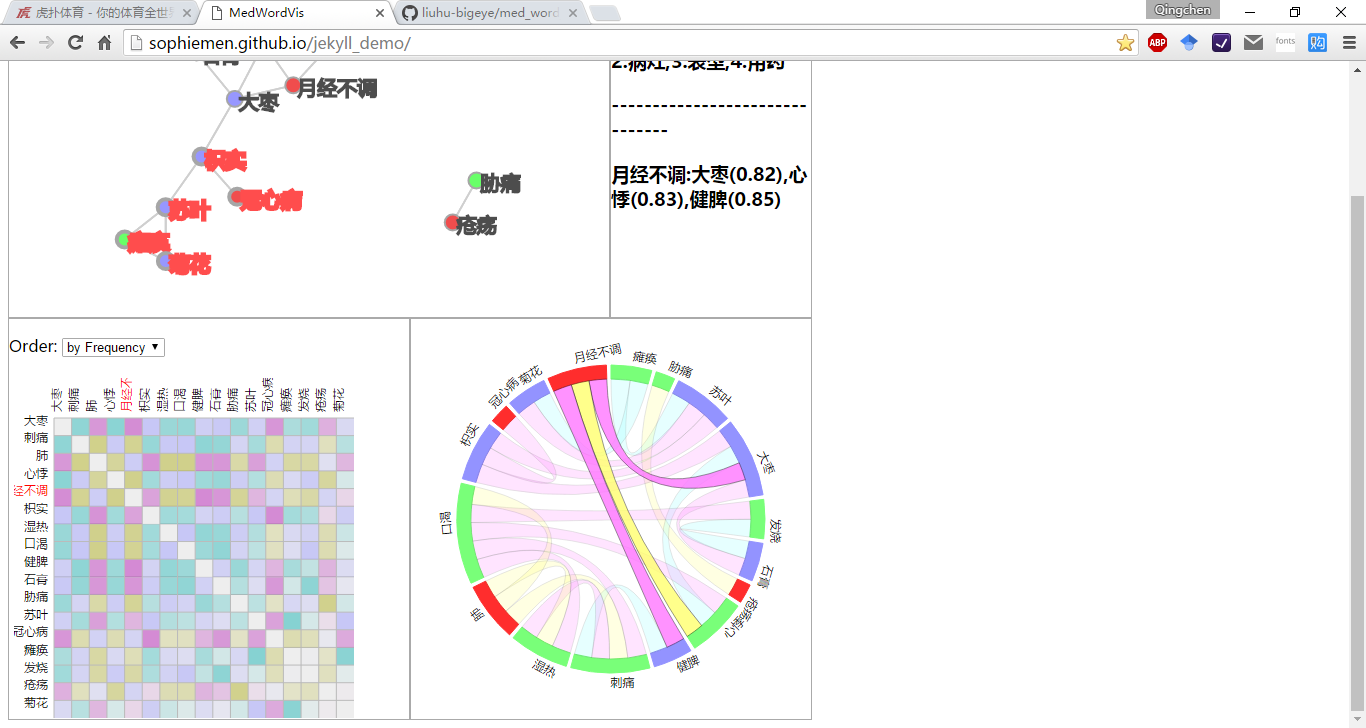


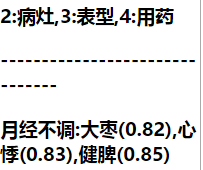
图 12 选中一个节点后信息显示

整体视图的动态调整和多视图的联动是我们这段时间调试的重点，最后的效果也比较好。

# 四、分析结果

我们的可视化不仅仅是一个可视化的窗口，目的是提供一个查询医学相关词汇的信息的窗口，也可以共专家进行验证和修改，相关的训练结果的准确度很大程度上取决于我们的训练语料的丰富和准确程度。和我们训练网络的迭代次数也有关系。

即使作为一个普通人，也可以从我们的结果中获得一些有用的医学信息，比如：月经不调词汇，和 大枣，心悸，健脾 有较强的关系，如下图：



当然，专家也可以通过网络的直观表达来探究一些过去被忽视的关系，比如枳实 和 冠心病，就有相关医学论文研究这二者的关系。

可视化的过程中，不仅仅需要将数据进行表达，而需要考虑受众的信息接收和互动交流，这里我们的项目是做了一些有益的尝试的。

我们的项目在github上有展示demo: <http://sophiemen.github.io/jekyll_demo/> 可以访问。

# 五、总结

数据可视化就要结束了，给我最大的收货就是相较于传统计算机学科比较重视数据的处理和分析，轻视可视化的表达和互动。其实数据可视化是一门很重要的学科，如何将信息更好的传达，如何形成良性的交互，是数据可视化的初衷。这方面涉及到很多数据挖掘，算法设计，美术设计等信息。感谢本课程给我们的成长和收获，本次项目不仅仅是学习一些新的工具，造新的轮子。如何面对具体问题和需求进行分析，使用适当方法，如何进行小组分工协作，都是本次课程项目的成长和收货。