

# 基于图论的文本分析

No Author Given

No Institute Given

**摘要** 本文通过有向多重图来描述文档的组织结构, 使用networkx工具进行图的统计分析, 得出'新型冠状病毒肺炎诊疗方案'1-7版的一些文本分析结论.

**Keywords:** 图论, networkx, 文本分析

## 1 方法

本文通过有向多重图描述文档结构: 非叶结点代表文档标题(包含一级标题到五级标题), 叶子结点代表标题下对正文的分词, 边的权重设为该边的入点的出度, 并令叶子结点的入边权重设置为1.

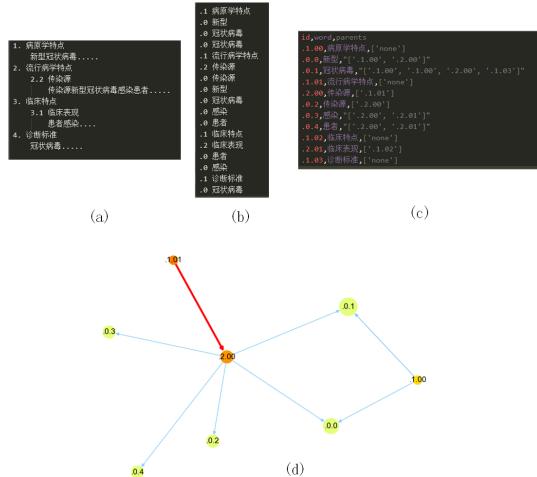
结点布局基于弹簧模型<sup>1</sup>, 如果两个结点父节点比较类似, 会自动放置在较近的位置.

本文从四个维度描述信息: 1)结点颜色, 令其正相关于结点后续结点个数, 个数越多颜色越深, 叶结点后续结点个数为0, 为绿色; 2)结点大小, 令其正相关于结点入度, 入度越大结点越大; 3)边颜色, 入点为非叶结点的边为红色, 为叶结点的为蓝色; 4)边粗细, 令其正相关于结点后续结点个数, 主干越粗后续结点越多.

为了处理文本数据, 首先将文档格式化, 结点标号, 然后读入组织成有向多重图, 如图1

为了分析文档关键词及其关系, 我们在第六版文档的基础上, 将所有结点画出, 并选择性滤除部分不重要结点, 得到了可以较好的表述文本关键词的词频和关系的图; 为了观察病毒特点, 症状, 诊断, 治疗正文内容的关系, 本文通过依次在图中添加其标题结点, 观察结构变化, 得出结论: 其四个部分内容公用大量关键词, 内容整体性角好; 为了分析文本随着版本的更替产生的变化规律, 抗击疫情过程中对病毒的认识的加深, 我们对不同段落章节单独分

<sup>1</sup> “spring” models (see Kamada and Kawai, Information Processing Letters 31:1, April 1989).



**图1.**一个例子. (a)格式化后的文档, (b)分词后文档, (c)有向图的存储文件, (d)最后组织出的有向多重图. 叶结点的入边为蓝色, 非叶结点入边为红色. 非叶结点颜色代表后续结点(子节点及其后代节点)数量, 叶结点的大小代表入度的大小.

表 1. 符号

标题	id
thr	如果叶结点入度小于thr则不显示
$\Omega$	根节点全集,包含 $1.00 \sim 1.11$
$G_\Omega$	以根节点为点集 $\Omega$ 张成的图
A	在图中显示的根节点集合

析,发现多数章节随着时间不仅内容变多,文档的分支也变得越来越多,反映出我们的诊断更为准确,治疗更有针对性.

## 2 关键词分析

通过有向多重图描述文档,非叶结点为文档标题,叶结点为正文分词.本次实验通过第六版进行测试,结构布局为”sfdp”<sup>2</sup>,文档第六版一共被格式化为11个一级标题,如表2

表 2. 一级标题

标题	id
病原学特点	.1.00
流行病学特点	.1.01
临床特点	.1.02
诊断标准	.1.03
临床分型	.1.04
鉴别诊断	.1.05
病例的发现与报告	.1.06
治疗	.1.07
中医治疗	.1.08

### 2.1 目的

1)将标题逐个添加分别得到子图,观察图的连接情况 2)分析关键词之间的关系,观察叶结点位置.根据结点的弹簧布局模型,得知若两个叶结点摆放位置相近,则其父节点类似,即代表的关键词常常一起出现在正文,关键词之间有高度的联系 3)观察图中叶结点的大小,越大说明结点代表的关键词频次越高,以此来分析结点频次

### 2.2 过程

令 $thr = 1$ ,根据表2中所有根节点,作出拓扑图2,来以此分析关键词关系.作图后发现叶子结点过多,不利于直观分析.经过观察发现由于中医专用

<sup>2</sup> 一种弹簧拉扯模型,使得总边长尽量小的同时,点尽量分开<https://www.graphviz.org/>



图 2.  $thr = 1, G_{\Omega}$ ,做出拓扑图,图中结点过多,不易分析

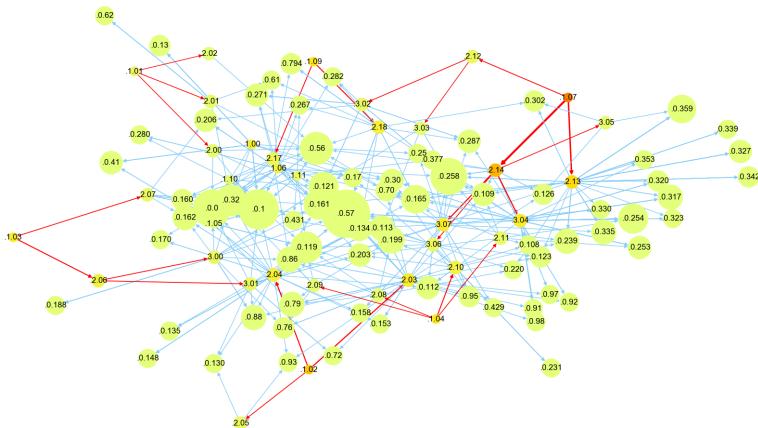


图 3.  $thr = 3, A = \Omega - .1.07$

名词与现代医学区别较大, 中医治疗这部分结点过多, 且与文本其他部分内容重用率低, 故不考虑, 去除结点1.08. 另外根据前文的词频统计结果, 发现频次为1的叶子结点占比过高, 为67.5%, 关键词频次低且不具有一般性, 故令 $thr = 3$ 滤除这部分结点. 过滤后得到如图3

## 2.3 结果分析

表 3. 词频较高的几个分如表所示,附件里有完整数据

id	word	频次
.0.619	10g	31
.0.57	患者	30
.0.511	15g	30
.0.697	注射液	26
.0.500	9g	21
.0.1	冠状病毒	18
.0.502	6g	17
.0.484	推荐	16
.0.0	新型	16
.0.335	每日	15
.0.258	治疗	14
.0.56	感染	13
.0.204	临床表现	13
.0.119	肺炎	11
.0.552	处方	11

根据图中结点的大小和边的粗细, 可以比较直观的看出文章结构的中心与高频关键词.

## 3 关键结点之间的关系

### 3.1 目的

为了明确”病毒特点——症状——诊断——治疗”之间的关系, 我们选取格式化后的几个重要标题作为根节点, 然后作图. 为了量化四个部分的关联程度, 我们定义, 在不同章节(结点)之间出现的相同关键词越多, 关联程度越

大, 具体表现为在初始图中添加关键结点的过程中, 整个图中高入度结点的增多. 现在定义上下文关联度为图中变数比点数, 即  $R = \frac{G_{edges}}{G_{nodes}}$

### 3.2 过程

根据表2, 我们定义以病原学特点, 临床特点, 鉴别诊断, 治疗作为点集  $A = \{.1.00, .1.02, .1.05, .1.07\}$  作为根节点, 以此作图4, 观察其过程.

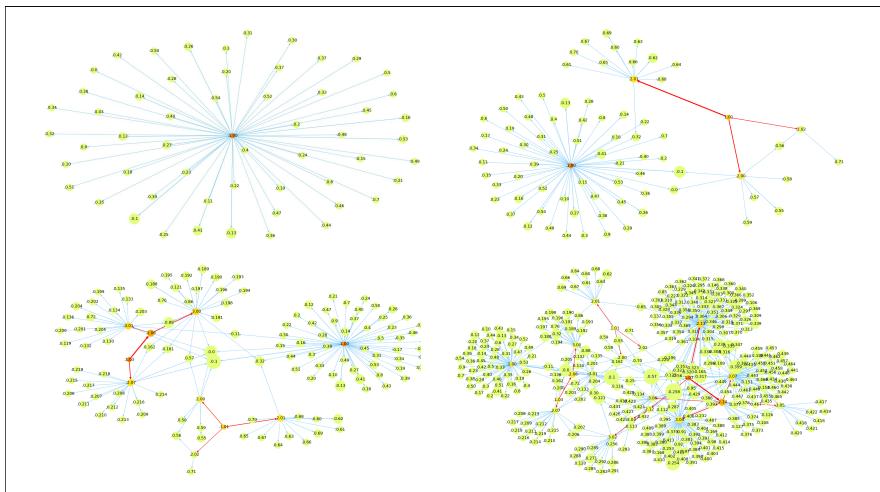


图 4. 在图中依次添加根结点的情况. 根节点为  $A = .1.00, .1.01, .1.03, .1.07,$

### 3.3 分析

根节点.1.00 为病毒病原学特点, 关键词多是病原学术语, 所以在添加结点.1.01“临床特点”时, 仅有.0.0“新型”, .0.1“冠状病毒”和.0.32“呼吸道”与其产生弱连通边; 添加结点.1.03“诊断标准”, 其含有两个二级节点两个三级结点, 其中.0.0(新型), .0.1(冠状病毒), .0.57(患者), .0.101(休克), .0.161(检测), .0.162(核酸), .0.11(特征), .0.32(呼吸道), .0.88(发病)产生关联. 添加.1.07(治疗)时, 产生关联较多. 另外再添加.1.08(中医治疗)后, 关键词数量激增, 关联度也明显增高, 侧面反映出说明中医治疗用的词汇与之前的区别较大, 共用词汇较少, 而且局部重用率较高.

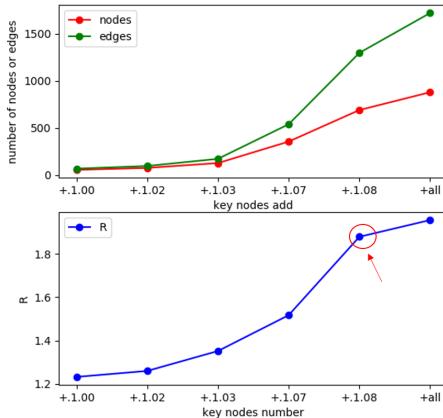


图 5. 在图中依次添加根结点时候图的连通度变化, 可见根据根节点的加入, 连通度在增加, 尤其是加入中医治疗结点后(红色箭头)

## 4 版本变化

### 4.1 目的

本段对新旧版本之间的文本做统计分析, 旨在寻找版本更迭的规律, 着重反应在文档的内容和结构上.

### 4.2 过程

本文对所有关键结点统计分类, 一共有13个一级节点, 对相应结点的不同版本进行统计, 得到如图6. 另外根据个别关键结点, 比如治疗和中医治疗, 为了比较直观的看出随着版本更迭其拓扑结构的变化, 我们将这部分各个版本的文档作图, 如从而能一定程度反映出

### 4.3 结果分析

1)根据图6, 容易看出基本在所有章节都会随着版本更迭, 关键词增多, 篇幅增大, 尤其是在临床特点, 治疗和中医治疗部分. 可以容易得知随着时间更替, 1)可以根据图8, 7看到, 根据版本更迭, 文档篇幅越来越多, 分支也变多, 结构变得更加详细, 体现出治疗更有针对性. 3)另外, 我们对“中医治疗”进行



图 6. 各个根节点的后继结点数量和边数量, 这能直接反应量化后的内容大小

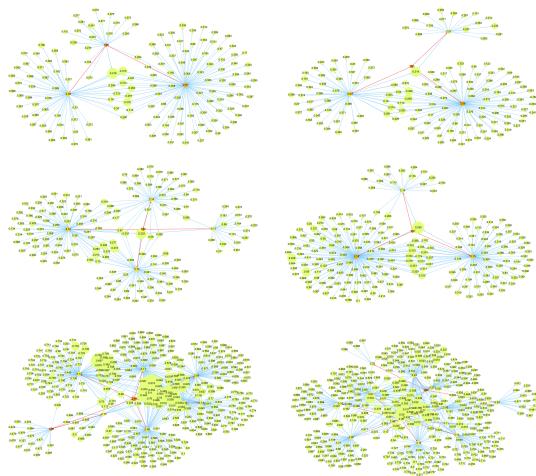


图 7. ”治疗”章节不同版本的文档结构变化.从左上到右下分别是第二版到第七版的拓扑图.

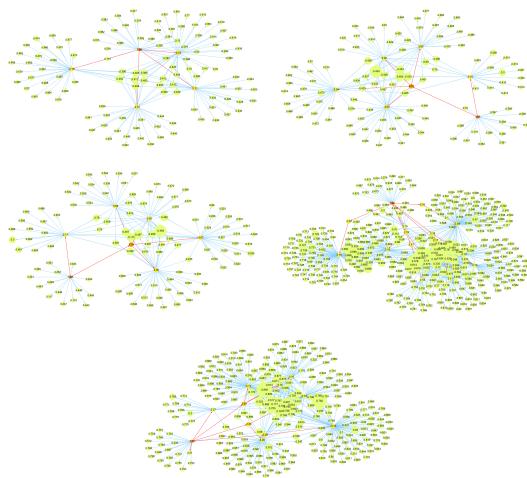


图 8. ”中医治疗”章节不同版本的文档结构变化.从左上到右下分别是第三版到第七版的拓扑图, 第二版及其之前并没提出中医在此疫情中的作用.

分析, 将干扰词汇比如计量用词, 冲服方法等滤除, 发现局部高频词中, 对于某些药材, 比如藿香与生石膏, 不仅频次高, 且在多个中药配方中一起出现.

## 5 总结

本文通过图论模型, 对新型冠状病毒肺炎诊疗方案1-7版进行可视化分析. 将文档标题抽象成非叶结点, 并将正文部分分词, 抽象为叶结点, 通过图论的统计分析工具, 得到了一些有用的结论, 其中包括: 1)所有版次章节的词频统计分析及其引用图; 2)文档病毒特点-症状-诊断-治疗几个章节的分词关系; 3)随着版本更替, 手册篇幅更长, 治疗手段更详细, 对病毒的理解渐渐加深.

10 No Author Given

## 参考文献