DOI:10.3969/j.issn.1005-202X.2019.05.024

其他

基于共词分析和可视化的高血压疾病关联性挖掘

刘莉[1],姚京京[1],李俊[2],陈先来[3],周宇葵[1]
1.中南大学生命科学学院,湖南 长沙 410013;2.中南大学湘雅口腔医学院,湖南 长沙 410008;3.中南大学信息安全与大数据研究院,湖南 长沙 410083

【摘要】**目的:**对高血压患者电子病历病案首页进行分析挖掘,揭示其中疾病诊断之间的关系。**方法:**以共词分析为基础,通过Python语言构建分析模块,采用Gephi复杂网络分析软件对结果进行展示。**结果:**基于3 632条电子病历记录,构建包含疾病诊断节点1 029个,共现关系边12 479条的疾病诊断共现网络,发现共现关系较强的疾病诊断集群。**结论:**从多角度、多层面对疾病诊断共现网络进行解读,并以可视化图谱的方式展示,揭示疾病诊断之间关系,为下一步构建更加完善的疾病图谱奠定基础。

【关键词】电子病历;病案首页;高血压;共词分析;可视化
【中图分类号】TP391;R312 【文献标志码】A 【文章编号】1005-202X(2019)05-0614-07

Hypertension related association mining based on co-word analysis and visualization

LIU Li[1], YAO Jingjing[1], LI Jun[2], CHEN Xianlai[3], ZHOU Yukui[1]
1. School of Life Science, Central South University, Changsha 410013, China; 2. Xiangya School of Stomatology, Central South University, Changsha 410008, China; 3. Institute of Information Security and Big Data, Central South University, Changsha 410083, China

Abstract: Objective To analyze and mine the electronic medical record home page of patients with hypertension, and to reveal the relationships among disease diagnoses. **Methods** Based on the co-word analysis, the analysis module was established by Python language, and the analysis results were displayed by Gephi complex network analysis software. **Results** Based on 3 362 electronic medical records, a disease diagnosis co-occurrence network containing 1 029 disease diagnosis nodes and 12 479 co-occurrence relationship lines was constructed, and a disease diagnosis cluster with strong co-occurrence relationship was found. **Conclusion** The disease diagnosis co-occurrence network can be interpreted from multi-angle and multi-level, and can be displayed in a visual map to reveal the relationships among disease diagnoses, laying a foundation for the further establishment of a more complete disease map.

Keywords: electronic medical record; medical record home page; hypertension; co-word analysis; visualization

前 言

高血压是一种以体循环动脉压升高为主要特征,遗传易感性和环境因素相互作用导致的全身性疾病[1]。已有大量研究表明有效控制高血压发病情况,可降低心脑血管疾病的患病风险[2-3]。电子病历(Electronic Medical Record, EMR)是病人的所有健康保健数据、病史及患病情况的存储[4]。作为一种新颖而丰富的临床研究资源,其研究价值不言而喻[5-6]。通过有效的数据可视化技术,电子病历数据中疾病诊断之间的关系可以以图形网络的形式清晰展示出来,以便医生探索其中的医学规则,辅助其进行疾病诊断,也可为患者提供直观的疾病关系网络。本研究以高血压相关的电子病历数据作为数据源,采用Gephi复杂网络分析软件和共词分析方法,从多角度、多层面分析展示病案首页中高血压相关诊断之间的关系,旨在揭示这些疾病诊断之间的联系,为下一步建立更加完善的疾病图谱奠定基础。

【收稿日期】2019-01-19
【基金项目】国家重点研发计划(2016YFC0901705)
【作者简介】刘莉,副教授,E-mail: 332140915@qq.com;姚京京,在读硕士研究生,E-mail: 446124548@qq.com
【通信作者】李俊,讲师,主要研究方向:医院信息系统、计算机网络,E-mail: lijun2016@csu.edu.cn

1 国内外相关研究

高血压及其相关疾病关系可以通过查阅文献资料、询问医疗工作者和访问医学网站等方式获取,其

中对电子病历进行分析是一种以患者为中心的研究方法，是了解患者患病情况的重要手段[7]。已有不少研究以电子病历为研究对象，开展自然语言处理、知识提取、可视化研究等方面的工作[8-10]。姚旭升等[11]以住院病案首页数据为研究对象，采用基于Apriori算法的关联规则挖掘数据流，建立疾病间关联规则模型。基于电子病历的分析可以发现患者最直接的信息，分析其中的规律，揭示各疾病之间的关系。

近年来，大数据的兴起和相关技术的迅速发展让生物医学成为发展最为迅速的领域之一[12]。在临床、药品、检验、影像和医学科研领域每天都产生着大量数据，并近乎以指数方式增长。因此，对这些医学领域的信息进行科学的收集、加工、分析、处理、展示，使其更好地为人类服务也就显得更加重要。基于共词分析构建共现网络的可视化技术探索关键词之间的关系并不是一项新的尝试，在许多领域都被有效利用，如研究文本分类中词的共现关系[13]，学科知识结构、研究热点分析[14-15]。共词分析用于确定各关键词之间共同出现的频次，使密切相关的关键词聚类，其可发现研究对象之间的关系和揭示潜在的可能关系[16]。

在高血压的研究领域中，多为临床研究、基础医学研究和数据挖掘研究，其中数据挖掘研究多集中于高血压识别模型和高血压症状研究，鲜有共词分析的可视化技术分析高血压及其相关疾病关系的研究报道。本研究旨在采用共词分析的可视化技术对病案首页诊断数据进行分析，构建高血压及其相关疾病的关系网络，分析与高血压相关的主要疾病之间的关系，为提供直观的高血压疾病关联图谱、展示临床已知的疾病关联、揭示潜在的与高血压相关疾病、辅助医生诊断提供参考。

2 数据源和研究方法

2.1 数据源

本研究选取湘雅三医院2017年11月份出院患者的病案首页数据作为实验数据源，共计记录3 632条，字段232个。基于患者隐私保护，首先对记录中的患者身份信息进行剔除，仅为每条记录随机赋予唯一识别码，以保证隐私信息的安全。以“高血压”为检索词，选择诊断字段中包含“高血压”的记录作为研究对象，共计808条记录。对所选记录和字段进行评估、筛选、填充、删除等预处理，最终获得四类字段。同时，以实验数据中的第一条记录为例，展示各字段的内容，其中门诊诊断和主要诊断结果不一定相同。实验数据中平均每条记录包含5.5个非空诊断字段，所含字段数量范围为3~17个，各记录非空字段数目分布整体呈偏态分布，记录非空字段数主要集中于4~10。

2.2 数据预处理

在电子病历数据中，病案首页数据的结构化程度相对较高，类似患者主诉等自然语言为主的字段较少，多为类似诊断信息等结构化程度较高的字段，表达简洁准确。但依旧存在因表达标准化不够完善、录入人员操作失误等情况。

由于患者的“其它诊断”数量具有个体差异性，诊断字段数量不尽相同，所以在实验研究中对空字段不进行填充处理。针对表达主题相同，但表达方式不同的字段内容进行转换处理，以提高一定的数据标准化程度，如“高血压Ⅲ”和“高血压Ⅲ级”则将两者统一以“高血压Ⅲ”进行表示。在本研究中“高血压Ⅱ”、“高血压Ⅲ”分别对应Ⅱ级高血压和Ⅲ级高血压，而“高血压”则是患者是否患有高血压的判断结果，可能为任意一级高血压。此外，针对记录中出现一些症状类诊断及诊断结果过于粗略的字段进行了删除处理。

在数据处理的过程中，未对诊断结果进行主题词、上下位词的匹配和调整，因此，会出现“高血压”、“高血压Ⅱ”和“高血压Ⅲ”等相似诊断名称。这主要是考虑到虽然经过主题词的调整和上下位词的缩放可以减少节点数量，使共现网络更加清晰，但会损失原本的疾病诊断信息，降低共现图谱的精度。

2.3 构建共现矩阵

共词分析研究的基础是基于两个假设：(1)两个关键词在同一条记录中同时出现，表明其所代表的主题之间具有关联性；(2)为探讨关键词之间相似度的聚类共现研究，需与研究的主题和目的保持一致[7]。基于共词分析的研究思想，把原始记录转换为原始矩阵，对原始矩阵进行分析处理生成共现矩阵，为下一步研究提供数据支持。

以Python语言编写处理程序，提取出原始矩阵中的共现关系，即获取原始矩阵中每一行任意两个元素的构成的无序共现对，并记录各元素出现次数和无序共现对出现的次数，其中元素出现次数以表格形式保存，共现关系以共现矩阵的形式表达出来，共现矩阵如式(1)所示。

$$\begin{bmatrix} 0 & c_1 & c_2 & \cdots & c_i \\ c_1 & 0 & v_{21} & \cdots & v_{i1} \\ c_2 & v_{21} & 0 & \cdots & v_{i2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_i & v_{1i} & v_{2i} & \cdots & 0 \end{bmatrix} \tag{1}$$

在式(1)中，c_i代表第i个关键词，v_{ab}代表第a个

关键词与第b个关键词的共现值，即两者同时出现在同一条记录中的次数。其中同一关键词之间不存在共现关系，其值为空，以0表示。据此所生成共现矩阵包含了原始矩阵中的共现关系和各关键词之间共现的强弱程度。

2.4 基于Gephi进行可视化分析和展示

Gephi是一款用于数据分析和复杂网络展示的免费开源工具，与用户有着良好的交互，可通过调整网络的布局、形状、颜色来显示隐藏的关系。本研究以病案首页诊断信息为节点，诊断间的共现关系为边，构建基于病案首页的高血压诊断相关共现图谱，借助Gephi软件的数据分析工具，从模块化、平均度、平均聚类系数等指标角度分析共现图谱，解读高血压诊断之间的相关关系。

在整个共现网络中，连接较为紧密的节点群可以被看成是一个社区，或划分为一个社区。模块度是评价社区划分优劣的重要指标，模块度的值越大，社区划分的效果越好，其简化公式如式(2)所示。

$$Q = O\sum_{c}\left[\frac{\sum in}{2m} - \left(\frac{\sum tot}{2m}\right)^{2}\right] \tag{2}$$

其中，$\sum in$表示社区c内部的权重，$\sum tot$表示与社区c内节点连接的边的权重，包括社区内部的边和社区外部的边。Gephi软件中的模块化计算采用Fast Unfolding算法，这一算法是为了寻求最大模块度值以达到最佳的社区划分结果[17]。疾病诊断共现网络通过模块化计算可得到多个关系较为密切的社区，便于进一步分析其中的关系。

在宏观层面上，主要以平均聚类系数对网络进行分析[18]。平均聚类系数是整个网络上节点倾向形成聚类程度的平均值，每个节点的聚类系数都在0~1的范围。若任一节点的聚类系数为0，表明该节点为独立节点，即没有其他节点与之相连，但本文仅提取了存在共现关系的疾病诊断信息进行研究，所以并不存在聚类系数为零的独立节点。若任一节点的聚类系数为1，则表明该节点与网络中所有节点都有直接或间接的相连关系，即存在路径连接任意节点。在疾病诊断共现网络中，平均聚类系数代表各诊断节点倾向于与其他节点共同出现的强度。

在微观层面上，主要以中介中心性(Betweenness Centrality)、接近中心性(Closeness Centrality)对网络进行分析[17]。中介中心性是指网络中经过某点并连接这两点的最短路径占这两点之间的最短路径线总数之比，强调该节点在其他节点之间的连接能力，可能是块之间的衔接桥梁。接近中心性是指每个结点到其它结点的最短路径之和的倒数，节点接近中心性的值越高，代表其在该网络中的中心位置，地位越重要。中介中心性和接近中心性相比，中介中心性强调的是节点在网络中的衔接桥梁作用，为整个网络的贡献程度，接近中心性更加强调节点自身的中心位置。

3 结果展示与分析

3.1 疾病诊断共现图谱整体分析

本实验数据共计808条记录，各记录非空字段总计18 997条，涉及疾病诊断结论1 029个，共现关系12 479条。其中，频次前10的疾病诊断名称如图1所示，可见这10个疾病诊断名称都是临床上普遍认可的高血压相关诊断，如2.2所述，未对疾病诊断结果进行主题词、上下位词的匹配和调整，导致出现“高血压”、“高血压Ⅲ”和“高血压Ⅱ”等相似诊断名称，以保证疾病诊断共现图谱的精度。

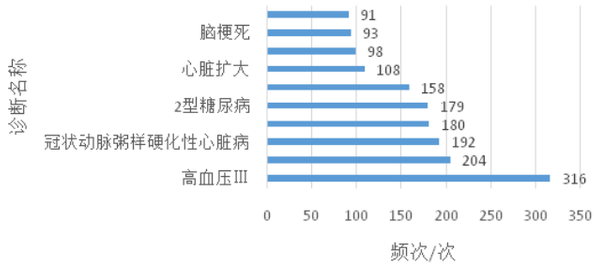


图1 频次前10的疾病诊断名称

Fig.1 Top 10 disease diagnoses

将1 029个诊断节点数据和12 479条共现关系边数据导入Gephi复杂网络分析软件，对其进行模块化分析，解析度设为默认值1.0，寻求最佳的社区分组。对模块化分析结果进行统计，共得社区分组11个，社区分组内节点占总节点数百分比较高的为社区分组1(36.73%)，社区分组2(26.53%)和社区分组3(14.97%)，该三大社区覆盖共现网络中78.23%的节点。其中所占比例超过10%的相对较大社区仅为3个，在后续社区分析中，将以这3个社区为研究对象。为全方面了解共现网络中的相关信息，对整个网络的信息进行统计分析，结果如表1所示。本节将从宏观和微观两个层面，基于共现网络指标数据对共现网络进行分析解读。

经过Gephi软件“模块化运算”后，并对同一社区设定唯一颜色。其中节点占比在1%以上的社区共有7个，分别对应的颜色为1(红)、2(绿)、3(深蓝)、4(淡蓝)、5(棕)、6(粉)、7(橙)。在图2中，展示了基于度和社区分

表1 诊断共现网络相关指标

Tab.1 Diagnosis co-occurrence network related indicators

指标	统计结果
诊断节点数	1 029
边数	12 479
平均度	24.255
网络直径	5
模块化	0.269
平均聚类系数	0.789
平均路径长度	2.351

组调整节点大小和颜色的疾病诊断共现图谱。从图2中可以清楚看出,其构图十分复杂,但仍可看到“高血压Ⅲ”、“高血压”、“高血压Ⅱ”、“2型糖尿病”等疾病诊断名称是关系图谱中的核心连接枢纽,其节点度数相对较大,也就是高共现的疾病诊断。

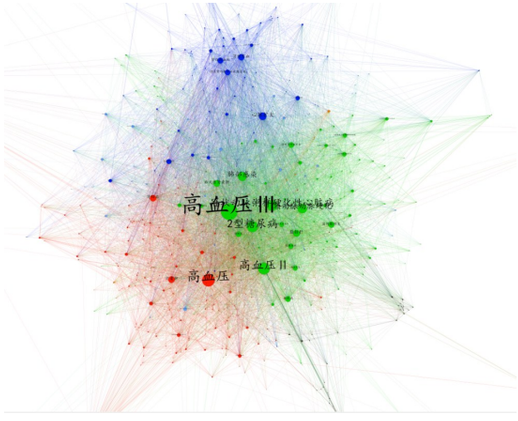


图2 高血压相关诊断共现图谱

Fig.2 Co-occurrence map of hypertension-related diagnoses

聚类系数是衡量网络中节点倾向于形成聚类的程度,聚类系数的高低意味着该节点所代表的诊断结果倾向于与其它诊断结果同时出现的程度。疾病贡献网络中聚类系数为1.0的节点总数较多,达到了526个节点,占总节点数的51.12%,代表半数左右的诊断倾向于与其它诊断同时出现的程度较高,其与相邻节点完全连接。不存在聚类系数为零的诊断节点,即不存在完全独立的诊断节点。其余部分疾病诊断节点聚类系数较为均匀的分布在0到1之间。因此,大部分诊断节点的聚类系数较高,平均聚类系数为0.789,表明大部分的疾病诊断都是倾向于与其它疾病诊断共同发生的。

3.2 中介中心性和接近中心性分析

为了揭示单个节点的属性,需要从相对微观的角度对疾病诊断共现网络进行分析。关于节点中间度测量的指标较多,其中,中介中心性和接近中心性两个指标最为重要[18]。本节将从中介中心性和接近中心性两个角度对疾病诊断共现网络进行分析。

中介中心性衡量了一个节点作为媒介者的能力,具有高中介性的节点被认为是便于管理和重要的节点。因此,这些存在于多诊断最短路径上的诊断信息可以认为是衔接诊断社区分组的桥梁,导致多种疾病共同出现。各节点中介中心性如图3所示。可见高中介中心性诊断节点分布稀疏,数量较少,而低中介中心性节点分布密集,集中于0~20 000。其中7个疾病诊断节点具有高中介中心性,其值从21 944到106 490不等,对网络的影响相对较大,值由高到低分别为高血压Ⅲ、高血压、高血压Ⅱ、2型糖尿病、阑尾术后、冠状动脉粥样硬化性心脏病、颈动脉动脉硬化。

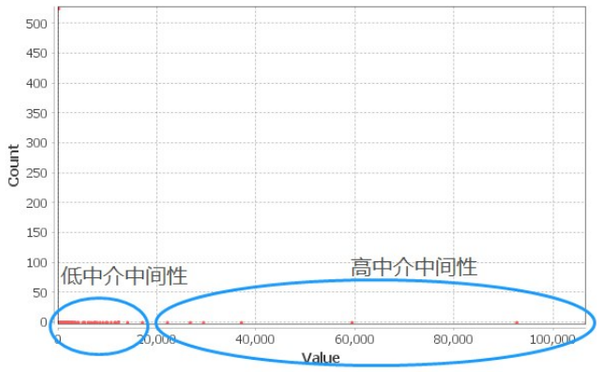


图3 诊断节点中介中心性分布

Fig.3 Betweenness centrality distribution of diagnostic nodes

接近中心性是从网络中的一个节点到所有其他节点的平均最短路径距离的度量。诊断节点的接近中心性越高,代表该节点处于网络中更加中心的位置,与其他诊断距离较近,关联性更强。高接近中心性的疾病诊断往往是临床上与高血压相关的常见病,可能是并发症、合并症等。诊断节点接近中心性分布图如图4所示,可见接近中心性分布较为均匀。“高血压Ⅲ”接近中心性最高,其后依次为高血压、2型糖尿病、高血压Ⅱ,与大部分节点接近中心性差距不大,节点整体分布较为连续,未出现集群分布。因此,疾病诊断共现网络,众多疾病诊断关系彼此之间相互交错,并没有疾病处于完全中心的地位。

如2.4所述,中介中心性强调节点在其他节点之间调节能力,控制能力指数,中介调节效应;而接近

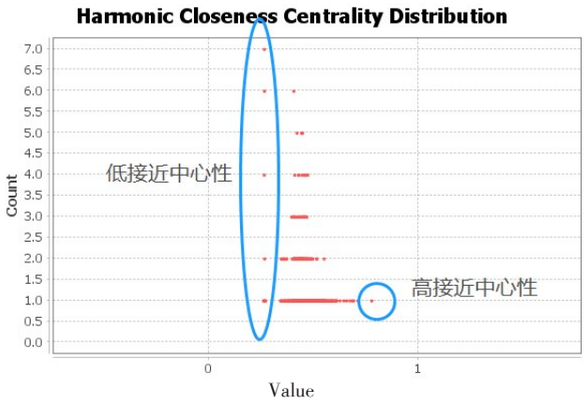


图4 诊断节点接近中心性分布

Fig.4 Closeness centrality distribution of diagnostic nodes

中心性强调节点在整个网络中的价值，价值越大，节点越处于中心位置。将节点中介中心性降序排列，分别以中介中心性和接近中心性为纵坐标构建折线图，以对比两者趋势变化，结果如图5所示。可见两者变化总体变化趋势相同，但彼此之间没有必然相关性，中介中心性越高，接近中心性不一定越高。

结合3.2和3.3的分析可知，“高血压Ⅲ”、“高血压”、“高血压Ⅱ”三者无论从平均度、平均聚类系数等宏观指标，还是中介中心性、接近中心性等微观指标来看，都处于疾病诊断共现网络中相对突出的的位置。同时，除聚类系数外，三者的度、中介中心性、接近中心性的值依次递减，“高血压Ⅲ”患者属于高

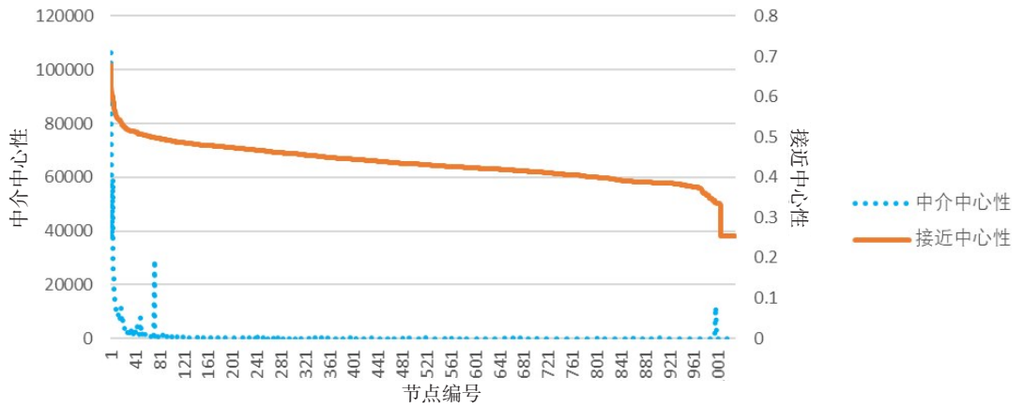


图5 中介中心性和接近中心性对比折线图

Fig.5 Line chart of betweenness centrality and closeness centrality

危人群，合并症或并发症更多，危及生命的风险更大，所以，住院比例相对更高。而“高血压”只是对患者是否患有高血压疾病的界定，其所占比例更大程度上是由医生选择基于“患者是否患有高血压”还是“患者所患高血压级别”下诊断结论所决定的。相比于“高血压Ⅲ”，“高血压Ⅱ”人群病情稍好，因而住院比例略微低一些。

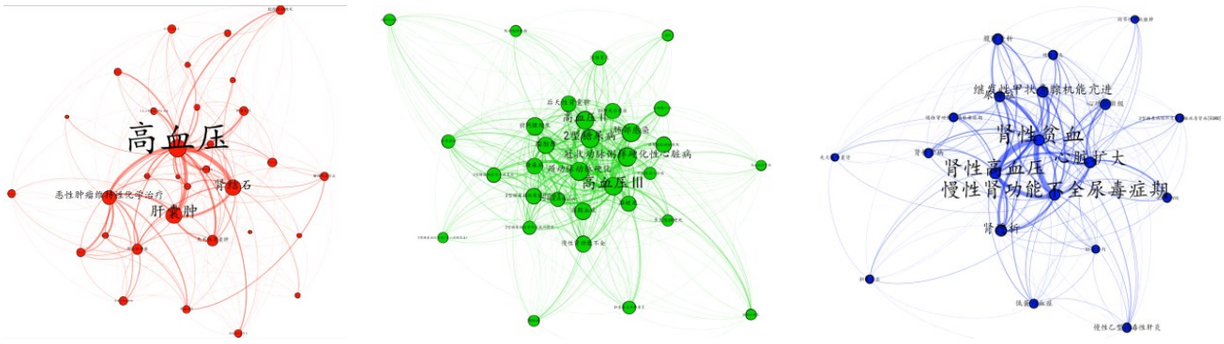
3.3 疾病诊断共现图谱社区分析

在3.1对疾病诊断共现网络模块化分析中，得到社区分组11个，但未对社区内节点内容进行分析研究，探讨各社区疾病诊断节点内容的关联性。本节对社区节点数排名前3且所占比例大于10%的3个社区进行研究。

图6a~c分别是社区1、社区2、社区3疾病诊断节点的关系网络，分别占总节点数的36.73%、26.53%、14.97%。由于社区内节点仍然较多，现过滤掉社区中度数相对较低的诊断节点，使图像更加清晰，便于展示分析。

在社区1中，高血压、肝囊肿、肾结石、先天性肾囊肿、恶性肿瘤维持性化学治疗度数最高，且从边的粗细可以看出彼此之间共现次数较高，在社区中无论是接近中心性还是中介中心性都相对较高，处于社区核心地位。可见高血压、肝囊肿、肾结石、先天性肾囊肿之间共现关系较为密切，但目前临床上仅认为上述4种疾病处于合并症的关系，彼此之间的作用机制尚未查阅到相关文献资料，因此，上述四者的关系仍需进一步探究。

在社区2中，高血压、2型糖尿病、冠状动脉粥样硬化性心脏病、颈动脉动脉硬化等诊断节点的度数、中介中心性和接近中心性都较高，处于社区1的中心地位。高血压与动脉粥样硬化两种疾病互为因果，相互作用，两者常同时存在。高血压和糖尿病均为常见病，两者关系密切，患有其中一种疾病的患者会大大增加患有另一疾病的风险，同时动脉粥样硬化与糖尿病关联性也较强，糖尿病患者动脉粥样硬化的发病率较无糖尿病者高两倍。



a:社区1　b:社区2　c:社区3

图6 社区内诊断节点共现网络

Fig.6 Community-wide diagnostic node co-occurrence network

在社区3中，节点数量虽然达到总节点数的14.97%，但其处于中心位置节点的度数比社区1和社区2的要小，以心脏扩大、肾性贫血、肾性高血压、慢性肾功能不全尿毒症期为代表。该社区主要包括心脏功能异常、高血压、肾功能异常之间的关系。高血压可导致心脏扩大，造成心脏功能异常，与肾脏疾病更是互为因果，彼此都可引起或加重另一方的病情，肾脏调解水与钠的能力会影响血压，而高血压和动脉粥样硬化会导致流入肾脏的血液也会减少，导致肾脏病变，或是加速既有的损伤。

4 结　论

常规的共现模型十分的直接和成熟，在文本挖掘等多领域均被有效利用，面对医疗领域的问题，该方法表现得“预测”能力较弱，“提取整理”能力较强[19]。在共现图谱中表现的关联关系多为临床上所熟知，其主要作用是对病案首页数据的提取、整理、发现，辅助挖掘未知或者未确认关联关系，而其自身的数据挖掘能力较弱。本研究采用Gephi复杂网络分析软件对高血压相关疾病诊断进行提取整理分析，发现其与糖尿病、肾脏疾病、肝脏疾病、心脏疾病等共现关联性较强，可能与高血压导致心脏负荷大、血液供给不足等有关，其中一些疾病的发生存在集群现象，通过可视化图谱展示疾病之间的内部关系，有助于观察多疾病间的联系。

在本研究基础上，可以引入新的共现逻辑、关联逻辑和有效的电子病历记录相似度匹配算法，数据源更加多元化，包含基因、疾病、症状等多方面的研究数据，可以有效提高图谱的预测效果[20]。其中对非结构化数据进行自然语言处理，通过专业的术语词典过滤，提取出有效的命名实体，可极大丰富图谱的内容。

【参考文献】

[1] 北京万方数据股份有限公司. 临床诊疗知识库_原发性高血压[DB/OL]. [2018- 04- 28]. http://lczl.med.wanfangdata.com.cn/Home/DiseaseDetail?Id=JB25292.
Wanfang Data Co, Ltd. Clinical diagnosis and treatment knowledge base_Essential hypertension[DB/OL]. [2018-04-28]. http://lczl.med.wanfangdata.com.cn/Home/DiseaseDetail?Id=JB25292.

[2] 陈伟伟, 高润霖, 刘力生, 等.《中国心血管病报告2016》概要[J]. 中国循环杂志, 2017, 32(6): 521-528.
CHEN W W, GAO R L, LIU L S, et al. China cardiovascular disease report 2016 summary[J]. Chinese Circulation Journal, 2017, 32(6): 521-528.

[3] 陈伟伟, 高润霖, 刘力生, 等.《中国心血管病报告2017》概要[J]. 中国循环杂志, 2018, 33(1): 1-8.
CHEN W W, GAO R L, LIU L S, et al. China cardiovascular disease report 2017 summary[J]. Chinese Circulation Journal, 2018, 33(1): 1-8.

[4] HANNAN T J. Electronic medical records[J]. Health Inf, 1996, 133: 133-148.

[5] KHO A N, PACHECO J A, PEISSIG P L, et al. Electronic medical records for genetic research: results of the eMERGE consortium[J]. Sci Trans Med, 2011, 79(3): 79re1.

[6] WASSERMAN R C. Electronic medical records (EMRs), epidemiology, and epistemology: reflections on EMRs and future pediatric clinical research[J]. Acad Pediatr, 2011, 11(4): 280-287.

[7] MENG X, YANG J J. Visual analysis for type 2 diabetes mellitus-based on electronic medical records[C]. International Conference on Smart Health. Springer International Publishing, 2014: 160-170.

[8] 谢剑, 周小茜, 童凌, 等. 基于中文分词的电子病历数据挖掘技术[J]. 湖南科技学院学报, 2016, 37(10): 54-59.
XIE J, ZHOU X Q, TONG L, et al. Electronic medical record data mining technology based on Chinese word segmentation[J]. Journal of Hunan University of Science and Engineering, 2016, 37(10): 54-59.

[9] 牟冬梅, 任珂. 三种数据挖掘算法在电子病历知识发现中的比较[J]. 现代图书情报技术, 2016, 32(6): 102-109.
MOU D M, REN K. Comparison of three kinds of data mining algorithms in knowledge discovery of electronic medical records[J]. Data Analysis and Knowledge Discovery, 2016, 32(6): 102-109.

[10] 王昱. 基于电子病历数据的临床决策支持研究[D]. 杭州: 浙江大学, 2016.
WANG Y. Research on clinical decision support based on electronic medical record data[D]. Hangzhou: Zhejiang University, 2016.

[11] 姚旭升, 杨静, 谢颖夫, 等. 关联规则算法在临床医疗诊断中的应用[J]. 软件导刊, 2018, 17(3): 162-164.
YAO X S, YANG J, XIE Y F, et al. Application of association rule algorithm in clinical medical diagnosis[J]. Software Guide, 2018, 17 (3): 162-164.

[12] 张艳. 大数据背景下的生物医学信息处理[J]. 生命科学仪器, 2014, 12(5): 17-20.
ZHANG Y. The processing of biomedical information in the era of big data[J]. Life Science Instruments, 2014, 12(5): 17-20.

[13] 章舜仲. 文本分类中词共现关系的研究及其应用[D]. 南京: 南京理工大学, 2010.
ZHANG S Z. Research and application of co-occurrence of words in text classification[D]. Nanjing: Nanjing University of Science and Technology, 2010.

[14] 曹树金, 吴育冰, 韦景竹, 等. 知识图谱研究的脉络、流派与趋势—基于SSCI与CSSCI期刊论文的计量与可视化[J]. 中国图书馆学报, 2015, 41(5): 16-34.
CAO S J, WU Y B, WEI J Z, et al. History, schools and trend in knowledge map: investigation and visualization based on SSCI and CSSCI[J]. Journal of Library Science in China, 2015, 41(5): 16-34.

[15] 秦长江. 基于科学计量学共现分析法的中国农史学科知识图谱构建研究[D]. 南京: 南京农业大学, 2009.
QIN C J. Research on construction of knowledge map of agricultural history in China based on scientometrics co-occurrence analysis[D]. Nanjing: Nanjing Agricultural College, 2009.

[16] CALLON M. Pinpointing industrial invention: an exploration of quantitative methods for the analysis of patents[M]//Mapping the Dynamics of Science and Technology. London: Palgrave Macmillan, 1986.

[17] BLONDEL V D, GUILLAUME J L, LAMBIOTTE R, et al. Fast unfolding of communities in large networks[J]. J Stat Mech, 2008, 2008(10): 155-168.

[18] AGGRAWAL N, ARORA A. Visualization, analysis and structural pattern infusion of DBLP co-authorship network using Gephi[C]. International Conference on Next Generation Computing Technologies. IEEE, 2017: 494-500.

[19] 傅博泉. 基于文本挖掘的基因—疾病关联关系研究[D]. 广州: 华南理工大学, 2016.
FU B Q. Research on gene-disease association based on text mining [D]. Guangzhou: South China University of Technology, 2016.

[20] WANG X, GULBAHCE N, YU H. Network-based methods for human disease gene prediction[J]. Brief Funct Genomics, 2011, 10(5): 280-293.

（编辑：黄开颜）