信息可视化及可视分析在智慧医疗领域的应用

关键词：信息可视化　可视分析　智慧医疗

曹　楠
同济大学

医疗健康是与每一个人直接密切相关的重要科学领域。科学家在探索生命奥秘和疾病产生机理的过程中，一直重视对跨学科技术的运用。从基于虚拟现实技术的仿真手术到手术机器人，从医学成像技术到医学图像处理，从大数据分析到人工智能，越来越多的新兴科技被应用到医疗领域，也提高了患者在就诊治疗过程中的安全保障。

长期以来，科学可视化 (scientific visualization) 技术[1]在医疗领域一直扮演重要角色。无论是平面 X 光扫描，还是三维 CT 影像，都应用了科学可视化的相关技术。然而这些技术仍然局限于对具象数据（例如人体骨骼、器官组织结构等）的展现。随着互联网的普及和可穿戴设备的广泛应用，越来越多与医疗相关的抽象数据被采集了上来，对信息可视化 (information visualization) 技术[2]提出新的需求，主要包括：(1) 展现用户的个人健康信息，例如心跳、血压等状态；(2) 汇总并展现公众健康信息，例如禽流感的扩散趋势、不同地区的人民健康状况等；(3) 分析并展现临床电子病历记录 (Electronic Health Record, EHR) 中的规律与模式，例如疾病的演变过程以及诊疗方案的疗效等。前两类可视化应用一般面向不具备医疗知识的普通用户，因此往往采用传统直观的信息可视化形式，如柱状图、折线图等，便于用户理解与阅读。第三类应用主要面向医生等具有专业背景，需要对数据进行深入调查并做出职业判断的用户，因此更具针对性，其可视化及相关分析技术的设计也更具挑战性。

在过去近十年，研究人员针对第三类可视化应用设计并开发了一系列与智慧医疗相关的可视化技术，主要用于：(1) 挖掘、展现并预测电子病历记录数据中潜在的规律及风险；(2) 帮助医生针对病患特征进行病人群体的相似性分析 (cohort analysis)；(3) 展现大规模医疗知识图谱 (knowledge graph)。

针对电子病历记录的可视化及可视分析

电子病历记录了患者就诊以及用药治疗的完整过程。它从患者和医生两个角度分别刻画了疾病在不同人群中演变发展的过程以及治疗方案在不同人

[1] 科学可视化是一个跨学科研究与应用领域，主要利用计算机图形学来创建视觉图像，帮助人们理解科学技术概念或结果的那些错综复杂而又往往规模庞大的数字表现形式。

[2] 信息可视化是一个跨学科领域，通过利用图形图像方面的技术与方法，帮助人们理解和分析数据。与科学可视化相比，信息可视化则侧重于抽象数据集，如非结构化文本或者高维空间当中的点。

群中所带来的不同疗效。因此，对于电子病历数据的分析与可视化具有重大临床意义。

电子病历数据及其带来的挑战

电子病历数据往往以文本或者表格的形式存储于计算机系统中。它可以被进一步抽象并转化成由医疗事件构成的事件序列数据。序列中的每个项记录了电子病历中的一次“就诊”“诊断”“用药”“化验”“手术”的记录，或者“入院 / 出院”这样的行为，或者“康复 / 死亡”这样的结果，以及这些事件发生的时间。针对这样的数据类型，可视化的设计空间可以由时间和事件两个主要的信息维度构成。

- **时间信息维度**：包括事件发生的时间，先后顺序，事件之间的时间间隔，周期性规律等可供设计的主要元素。
- **事件信息维度**：包括单个事件的类型，出现的频率，对应参数属性的大小（例如化验结果、用药剂量等），以及多个事件同时出现（或共同发生）的规律 (co-occurrence) 等设计元素。

任何针对事件序列数据的可视化都可看作是在这两个信息维度上挑选不同的设计元素，采用不同的设计方案及编码方式而构成的。然而，这个看似简单的任务在实际应用中却面临着诸多挑战。

首先，就时间信息维度而言，同样的疾病，有可能因为患者的自身差异、不同的时间安排、医生采用不同的诊疗手段等诸多因素，导致电子病历所记录的事件序列具有极大的差异，很可能同一组事件在不同患者身上，发生的具体时间、先后顺序、持续的周期、对应的属性取值均不相同，为事件序列数据的汇总与可视化带来了困难。

其次，从事件信息维度而言，考虑到药物、化验以及治疗手段的多样性，真实的电子病历数据中可能包含有数以万计的事件类型（例如，服用不同的药物均可看作是不同的医疗事件），从而导致数据具有高维异构性。不仅如此，真实的电子病历数据往往规模庞大，包含数以万计的病人和历时数十年的记录。这些都为可视化的设计带来了挑战。

大规模电子病历数据的可视化汇总

对大规模电子病历数据（即事件序列）进行汇总，是挖掘数据中潜在模式的必要手段。现有基于数据挖掘算法的技术实现了对原始事件序列数据的高度概括。通过分析，能够直接获得频繁子序列等事件序列中的相关模式[1]，但同时也丢失了数据中的细节及上下文信息。而现有的可视化技术，例如 EventFlow[2] 和 CareFlow[3]，多是对原始数据的直接展现，无法对较大规模的事件序列进行汇总。因此需要一种既能够汇总展现大规模事件序列的数据信息，又能够展现足够上下文细节的可视化技术。

为此，研究人员提出了 EventThread[4]，一种面向大规模事件序列数据的可视化汇总方案。如图 1 所示，该技术通过一系列数据处理及分析步骤，将原始的事件序列汇总转换成用于可视化的“潜在序列 (latent threads)”。在具体应用中，潜在序列可被视为一种潜在的“治疗方案”。该技术首先剔除了数据中的小概率事件，以减小数据噪声（图 1(a)）。接着，将事件序列靠左对齐（图 1(b)），并将对齐后的序列按照定长的时间窗口切割成若干个阶段（图 1(c)），用于代表离散化的时间维度。基于上述处理，该技术进一步将数据转换成一个包含病患、

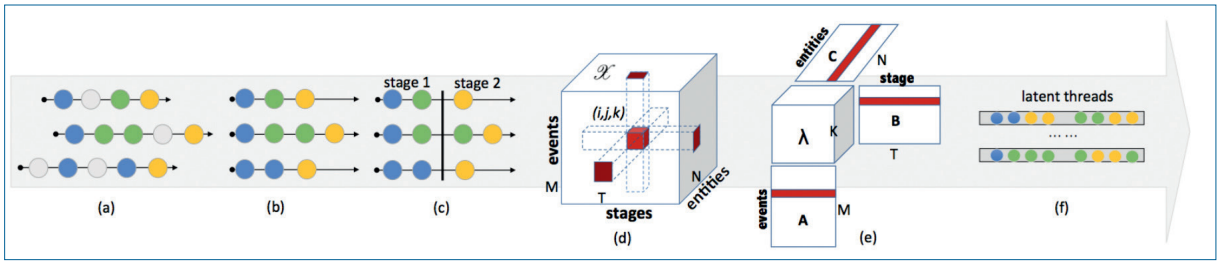

图 1　EventThread 对事件序列的处理及建模流程

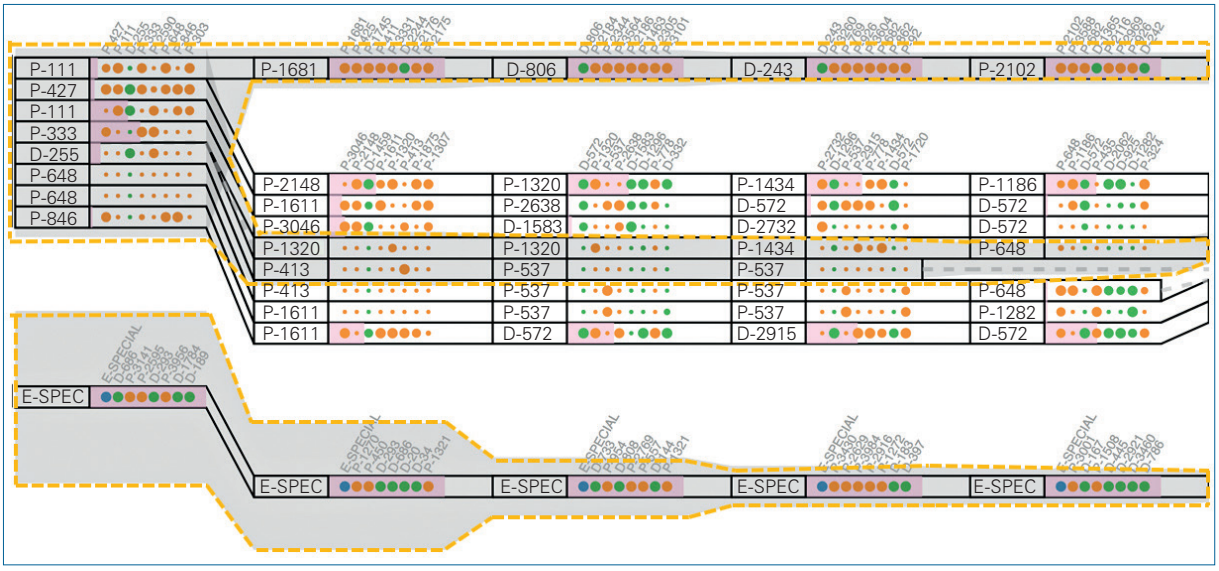

图2 EventThread 可视化设计。每一个“潜在序列”被画成平行的一行，相类似序列之间的距离也较为接近。序列中的小点代表对应阶段中的关键事件

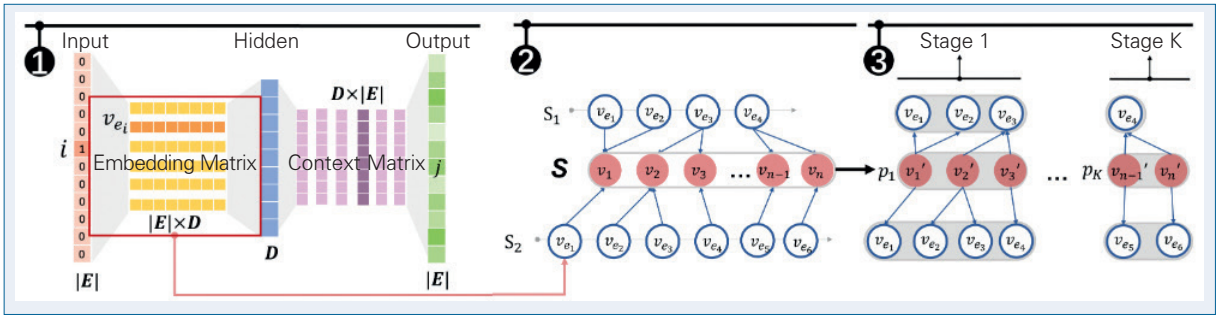

图 3 EventThread-2 中针对事件序列数据的阶段性分析

阶段、事件三个维度的张量（图 1(d)），并将其进一步分解为三个因素矩阵 (***A***, ***B***, ***C***)。其中，***B*** 和 ***C*** 分别代表了潜在序列（治疗方案）在病人群体中以及在时间维度之上的分布，而 A 则代表了事件在“潜在方案”上的分布，即该方案是由哪些具体医疗事件构成的。结合 ***A*** 和 ***B*** 中的信息，便能够得到“具体有哪些病人，在什么阶段与哪一个‘方案’相关”这样的上下文信息。

图 2 展现了我们对上述数据处理结果的可视化展现方案。该图汇总显示了 2 万名“慢阻肺”患者前后 8 年的电子病历数据。通过直接展现“潜在序列”，我们能够很好地汇总大规模电子病历数据。通过结合因素矩阵中的相关信息，我们能够为每一个序列提供足够的细节及上下文信息，以帮助用户理解每一个序列的具体含义。图 2 中的可视化设计清晰地揭示了“慢阻肺”采用的三种不同的诊疗方案，其中中间的一组是较为主流的方案，由多根并列的序列共同展示。

针对疾病演变的阶段性分析及可视化

EventThread 方案虽然能够汇总并显示大规模数据，但存在两点缺陷：(1) 未能解决序列多样性的问题，简单地将序列靠左对齐，并不是一个合理的方法；(2) 将序列分割成阶段时，使用了固定长度的事件窗口，这样的分割有可能把本来连续发生的事件序列（例如一次住院过程中所产生的所有事件），分

割成两个不同的阶段，从而导致分析错误。为了解决这两个问题，研究人员对 EventThread 方案做了进一步修改，提出了 EventThread-2[5]。该方案采用新的数据分析方法及可视化设计方案，针对事件序列数据进行阶段性分析，以更加清晰的视角展现疾病的演变过程或诊疗方案的阶段性进展。

图 3 展现了 EventThread-2 所提出的阶段性分析算法。该算法首先通过神经网络将事件序列数据中的每一个事件，根据其相互之间的相关程度映射到高维空间，实现事件到向量的转变（图 3 ①）。接着，通过动态时间规整 (dynamic time wrapping) 算法，将长短不一、顺序各异的事件序列根据相关事件之间的相似度进行对齐，并将对齐后的结果汇总在一个虚拟序列 S 之上（图 3 ②）。最后，通过对 S 进行切分，划分出序列发展的不同阶段。在此过程中，算法确保被划分在同一段中的事件具有较大的相似度，而不同段之间的事件在向量空间中具有较大的差异性，将切分结果进行拆分，可得到针对每一条序列的阶段性分析结果。该技术很好地解决了事件序列长短、顺序不一致等问题。

图 4 展示了利用 EventThread-2 技术对一组患有心脏病的 ICU 患者的电子病历数据及相应事件序列进行阶段性分析的结果，清晰地再现了患者从入院到化验，再到药物治疗、手术、及最终出院的全部治疗过程。可以看出，该技术能够准确地对医疗事件进行切分与汇总。

病人群体的相似度可视分析

病人之间的相似性分析在医疗领域也具有重大意义。首先，当遇到疑难杂症时，在病例库中查询相似病人及相应的诊疗方案对救治当前患者具有一定的参考意义。除此之外，将病人按照相似度进行归类，有助于医生根据不同的病患特征，制定不同的诊疗方案。现有的基于机器学习的相似度分析往往缺少可解释性，在给出病人之间相似度数值的同

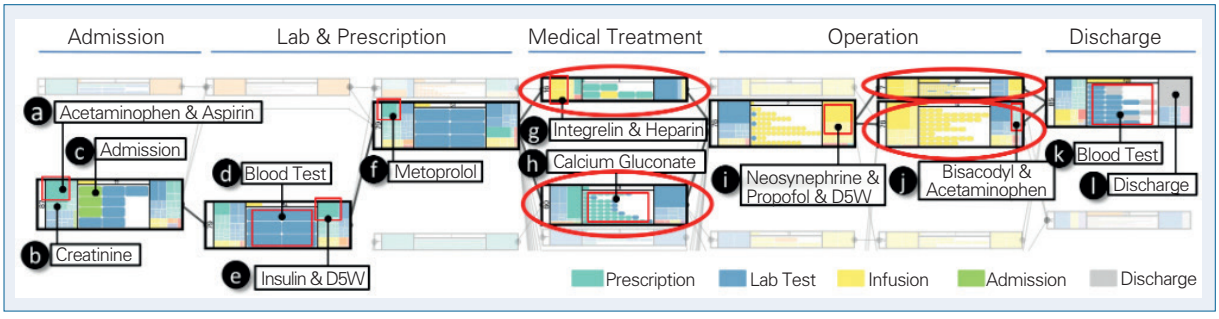

图 4　针对 ICU 患者入院序列的阶段性分析及可视化展示

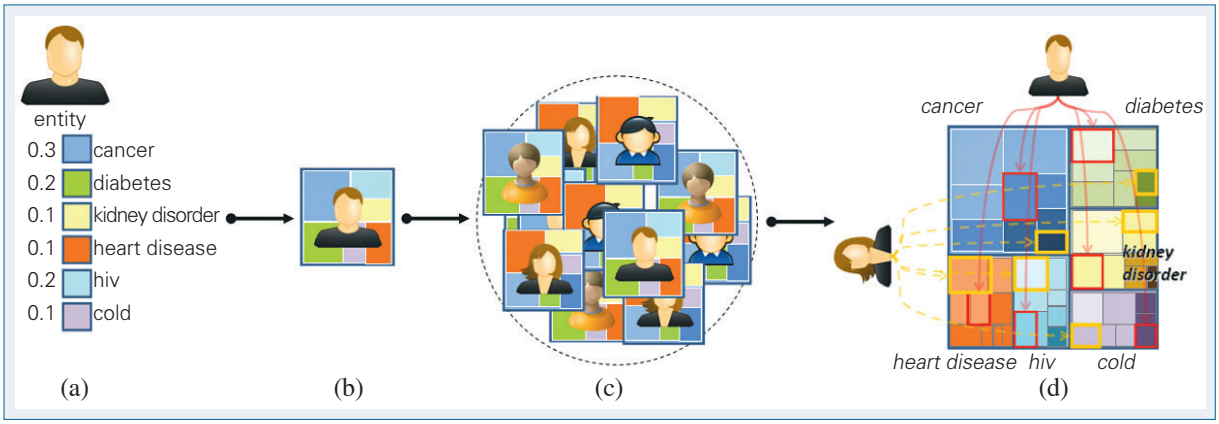

图 5　利用动态图标技术展示病人间的相似程度

时，不能够清晰直观地为缺少技术背景的医生解释病人为什么相似，在哪里相似。因此可视分析技术在该研究方向上的应用也非常重要。

为了提供直观的相似度分析工具，研究人员开发了动态图标技术 DICON[6]，用于直观展现病人的多维度特征，方便不同病患之间的相似度对比。该技术仍然基于对电子病历数据的分析，但在设计 DICON 时，我们忽略事件维度的信息，仅统计病人因特定疾病去就诊的次数，并以此作为病人的特征。如图 5 所示，根据电子病历记录，每一位病人都可能患有多种疾病，每一种疾病都被展示成一个矩形，疾病的种类用不同的颜色表示（图 5(a)）。病人因病就诊的次数被归一化处理后可以用矩形的大小表示，因此每一位病人都可以通过封装代表疾病的小方块，而被展示成一个方形的图标（图 5(b)）。面对一个拥有众多患者的病人群体，这样的设计依旧有效。通过拆分重组，我们把群体中不同病人的相同疾病封装在一起，从而构成了整个群体的图标（图 5(d)）。

这种可视化设计充分利用了图标技术的高可比性，让具有相似疾病的患者有相似的表示，不相似的患者有所区分。同时，我们对病患个体以及群体采用类似的设计方案以及完全一样的可视化编码方式，从而降低了可视化在理解上的难度。动态图标技术还可以与其他可视化图表一起使用。如图 6 展示了把动态图标技术应用在散点图中，用来展示多维度上下文信息的场景。

医疗知识图谱的构建及可视化

医疗知识图谱涵盖了医疗领域的相关概念（例如疾病、药物、症状等），主要用于构建智慧医疗体系中的自动问答系统。与其他知识图谱类似，医疗知识图谱往往是异构的，并且包涵大量节点与链接。知识图谱的构建需要对大规模的文本文件进行处理，以提取其中的相关概念，并构建这些概念之间的联系以构成知识本体。此外还需要直观展现知识图谱中所包含的复杂关系，提供高效的查询机制，从而提高分析人员浏览知识图谱的效率。

基于上述需求，研究人员设计开发了一系列针对文本数据的知识图谱构建及可视化系统。如图 7 所示，研究人员首先对疾病文档进行分析处理，将每一个记录疾病信息的文本文件按照疾病不同方面的描述（如症状、治疗方案、病因等）切割成多个信息层面，并从不同的信息层中提取实体关键词（如疾病或症状的名称），去重后构成实体关键词集合。基于该集合，实体之间分别以“内在关联”和“外在关联”两种形式建立相互联系。“内在关联”是实体在同一信息层面出现的关联（如两种互为并发

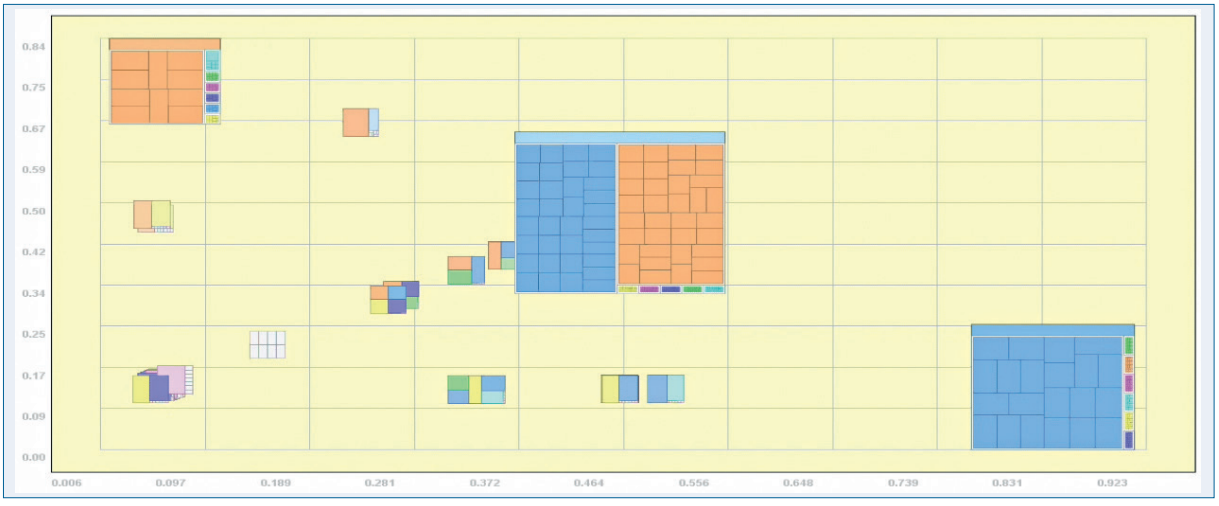

图6 在散点图中使用动态图标技术。横轴及纵轴分别代表肾功能失调（蓝色）及糖尿病（橙色）两种不同疾病，其他颜色代表其类型的疾病

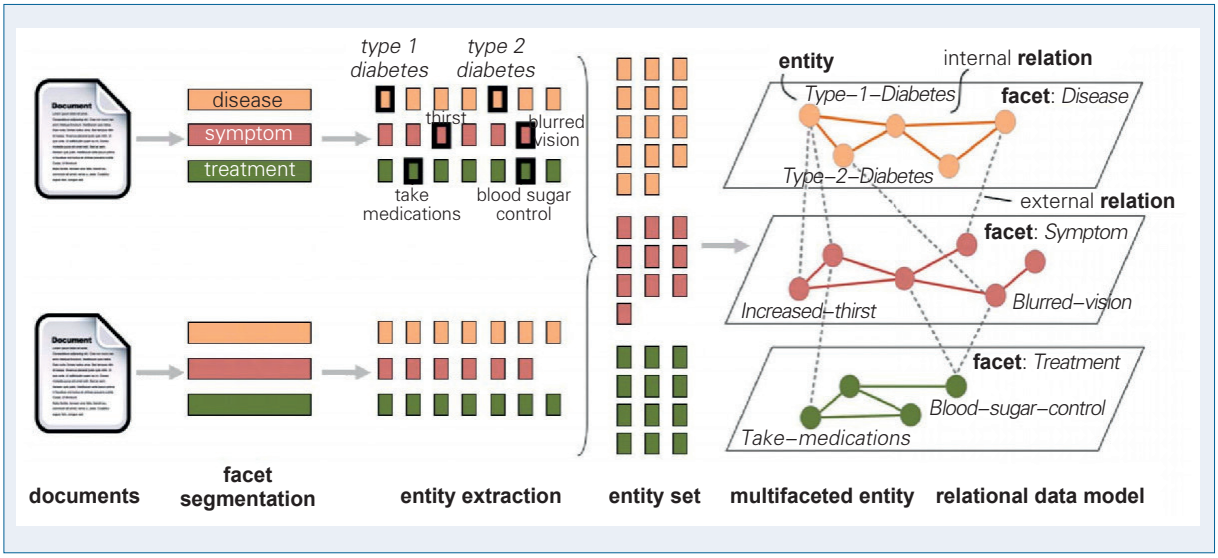

图 7　疾病知识图谱的构建以及多层次实体关系数据模型

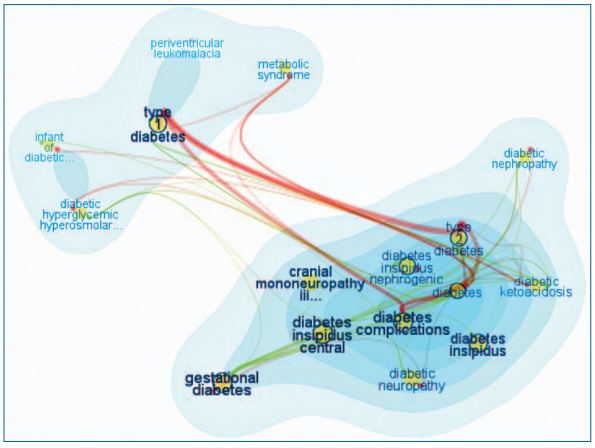

图8　利用 FaceAtlas 展现的两类糖尿病及其并发症之间在症状（红线）及治疗方案（绿线）之间的联系

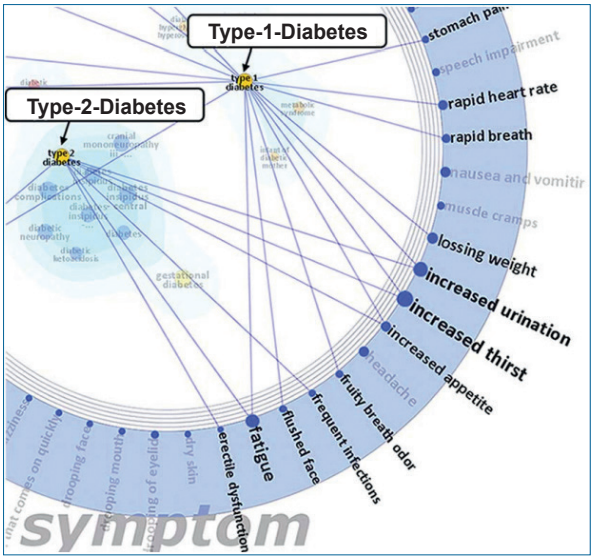

图 9　SolarMap 被设计用来展现不同实体之间的详细外在联系

症的疾病），而“外在关联”则是实体在不同信息层面出现的关联（如疾病及其症状之间的联系）。这样一系列处理将无结构的原始疾病信息文本文件转换成一个结构化的“多层面实体关系数据模型”，构成了医疗知识图谱中的内部基础结构。基于该模型，研究人员提出了多种相关的可视化设计，用于展现模型中蕴含的不同信息。

FaceAtlas[7] 是医疗知识图谱的一种可视化方案，该方案聚焦于显示同层面实体之间的外在联系。图 8 展现了两个分别以 1 型糖尿病和 2 型糖尿病为核心的类。红线代表了疾病之间在病因上的关联，而绿线代表了疾病之间在症状上的关联。FaceAtlas 还提供了完整的文本搜索机制，用户可以通过关键词搜索相关疾病，系统将自动构建与该疾病相关的知识图谱，并以可视化的形式展示。

虽然 FacetAltas 展示了同层信息之间在不同层

面上关联的强弱（例如，疾病之间在症状、病因、治疗方案上的联系分别用不同颜色的线表示，线的粗细代表了联系的强弱），但无法显示两个疾病具体在哪些症状上有关联。为了克服这一设计缺陷，研究人员又进一步设计了 SolarMap[8]。该可视化设计首先选择一个主要的信息层面，然后把该层面中的主要实体以图的形式显示在可视化视图的中心位置。如图 9 中所显示的可视化结果以疾病作为主要信息层面，可视化的中心视图展示了疾病及其并发症构成的网络。围绕着中心视图，其他信息层面被显示成一层层的圆环，用户可以通过交互的方式选择展开某一层次的圆环，相关信息层面中的实体信息以关键词的形式显示在该圆环的相应位置，这些实体与中心实体之间的关系则通过连线来展示。这样用户可以清晰地看到两种疾病在症状上的具体联系。

研究机遇与挑战

信息可视化及可视分析技术在智慧医疗领域的诸多方面都起着举足轻重的作用。然而，这方面的研究仍处于较为初级的阶段，有很多问题尚待解决，也存在诸多挑战与机遇。

第一，从数据角度而言，如何对大规模异构医疗数据进行融合以方便分析，仍然是一个较大的挑战。真实世界的医疗数据往往是非常复杂的，既包括像 X 光片、CT 扫描等各种影像数据，也包括化验结果等结构化的表格数据，还包含诊断、医嘱等无结构文本数据。医生需要结合上述所有信息方可对病人作出诊断，这对数据分析技术提出了挑战。如何融合并展现来自各种渠道不同类型的数据，并建立它们之间的关联，方便分析人员从不同视角进行观测和总结，并帮助他们汇总理解相应的分析结果，是可视化技术面临的一个重要挑战。

第二，从技术角度而言，随着人工智能技术在智慧医疗领域的广泛应用，越来越多的辅助诊疗系统及算法被开发出来，辅助判断患者可能的疾病状况，推荐相应的诊疗方案，以及对当前疾病进行预后分析。然而，所有的分析算法及人工智能技术均面临着可解释性的问题。当算法做出的判断无法解释或者不足以让医生信服时，将无法真正在医疗领域进行部署。因此有必要设计新的可视化技术，帮助全面解释算法分析的结果，回答诸如“为什么预测 A 疾病发生而不是 B 疾病？”“服用药物 A 的预期疗效为什么比服用药物 B 好？”等一系列问题，以帮助用户理解相应的计算结果。

第三，从医学角度而言，不同疾病拥有不同的病因、发病机理以及演变过程，往往无法用一种通用形式对所有类型的疾病进行分析展现，而需要针对不同的疾病设计不同的分析算法及可视化技术。相应地，可视化设计人员需要对相关医疗知识进行较为深入的理解与学习，从而大大延长了可视化设计开发的周期。因此，如何根据疾病的大类别汇总分析相关数据中的共有属性，在医学理论的基础上增加可视化设计的通用性，有待进一步研究。 ■

曹 楠
同济大学教授，博士生导师。主要研究方向为信息可视化及可视化分析。
nan.cao@tongji.edu.cn

参考文献

[1] Han J, Cheng H, Xin D, et al. Frequent pattern mining: current status and future directions[J]. *Data Mining and Knowledge Discovery*, 2007,15(1): 55-86.

[2] Monroe M, Lan R, Olmo J M, et al. The challenges of specifying intervals and absences in temporal queries: a graphical language approach[C]//*ACM Conference on Human Factors in Computing Systems*(CHI). ACM, 2013: 2349-2358.

[3] Wongsuphasawat K, Gotz D. Exploring Flow, Factors, and Outcomes of Temporal Event Sequences with the Outflow Visualization[J]. *IEEE Transactions on Visualization and Computer Graphics*, 2012,18(12): 2659-2668.

[4] Guo S, Xu K, Zhao R, et al. EventThread: Visual Summarization and Stage Analysis of Event Sequence Data[J]. *IEEE Transactions on Visualization and Computer Graphics*, 2018, 24(1): 56-65.

更多参考文献：http://dl.ccf.org.cn/cccf/list