文章编号：2095-2163(2019)02-0208-04 中图分类号：TP391；Q811.4 文献标志码：A

疾病本体数据在人体生理位置上的可视化研究

彭栩生，王亚东
（哈尔滨工业大学 计算机科学与技术学院，哈尔滨 150001）

摘 要：疾病本体是人类疾病的标准化本体，为了更加清晰地表示疾病本体并加快其检索速度，本文提出了一种将疾病本体数据映射到人体生理位置上的可视化方案。首先通过人工校对，将疾病本体数据与人体的系统或者器官对应，之后，从原始的 SVG 出发，通过对其分割之后再合并获得需要展示的人体系统或者器官的图片，并通过算法获取图片中内容的边缘路径，从而响应点击事件。结果表面，该可视化方案可以清楚且快速地对疾病本体进行检索。

关键词：疾病本体；可视化；生理

Research on visualization of disease ontology in human physiological position

PENG Xusheng, WANG Yadong
(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

【Abstract】 The disease ontology is the standardized ontology of human disease. In order to express disease ontology more clearly and speed up the retrieval speed, in this paper, a visualization scheme is proposed to map the disease ontology data to the physiological location of the human body. Firstly, the disease ontology data corresponds to the human body system or organ by artificial proof reading. After that, the original SVG is segmented and merged to obtain a picture of the human system or organ that needs to be displayed, and obtain the path of the image to respond to the click event. Results show that the visualization scheme can be used to retrieve the disease ontology clearly and quickly.

【Key words】 disease ontology; visualization; physiology

0 引 言

疾病本体[1-2]是人类疾病的标准化本体，对其开发将旨在为生物医学界提供一致的和可重复使用的人类疾病术语、表型特征和相关医学词汇。

疾病本体的拓扑结构是有向无环图，其根结点 id 为“DOID:4”，目前其下又可以分为 8 类，分别是：disease by infectious agent、disease of anatomical entity、disease of cellular proliferation、disease of mental health、disease of metabolism、genetic disease、physical disorder、syndrome 。

考虑到疾病本体数据繁多，且多以 obo[3-4]或者 owl[4-6]之类的文本形式存在，难以阅读和查找，因此需要对其进行合适的可视化操作。

本文给出了一种对部分疾病本体数据进行可视化的方案，可以将疾病本体数据映射到人体的生理位置，直观清晰地表示某个生理系统或者器官中的疾病，从而加速对疾病本体的筛选和查找过程。

1 总体方案设计

该可视化方案主要是在可视化区域中，首先绘制人体各主要系统或者器官的图片，另列出该层级下所有包含的疾病本体实例。当用户点击某个具体的器官或者系统时，可视化区域中所绘制的人体图片也将发生变化，仅绘制与选择器官或者系统有关的数据，与此同时，疾病本体列表也进行精简，只保留该系统或者器官下的实例。系统共分为 3 个层级，从上至下依次为：全身、系统、器官。对器官之下的、比如组织层级，由于图片数据等未臻细致，不再做进一步的区分，但本系统提供了后续的扩展性，若在将来当相应的数据映射和图片得到了补充，也可以给出更加细化的表示。

由前述总体方案设计可以得出，该可视化方案的设计流程可阐释为如下 3 个关键步骤：

(1)将疾病本体实例与某一系统或者器官进行映射。

(2)人体各系统和器官的图片的绘制。

(3)确定某一个点击事件所对应的系统或者器

作者简介：彭栩生(1992-)，男，硕士研究生，主要研究方向：生物信息；王亚东(1964-)，男，教授，博士生导师，主要研究方向：生物信息学、人工智能、知识工程等。

收稿日期：2017-06-15

官。

这里，在疾病本体中，“disease of anatomical entity”为解剖学实体相关的疾病，其 id 为 DOID:7，可以与人体的生理位置实现有效映射。因此，本次可视化方案中，选择疾病本体下的所有父亲节点或者祖先节点为 DOID:7 的子集作为疾病数据集合。该集合共有大约 3 000 条数据实例。这些疾病本体的实例，通过人工校对，被标注到对应的器官或者系统上。

2　关键技术设计和实现

2.1　人体各系统和器官的图片绘制

由本系统所使用图片文件来源于互联网，而且该图片已被授权可以用作包括商业在内的任意用途，同时使用者也可以自由更改。原始图片格式为 SVG[7-9]，设计研究内容如图 1 所示。

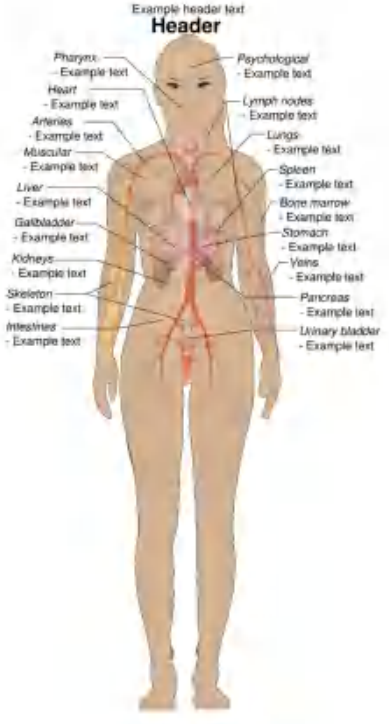

图 1　原始图片文件内容

Fig. 1　The original SVG file

在本质上，SVG 是 XML[10] 文件格式，可以通过其中文本的操作较为简捷地改变文件的内容。由图 1 可以看到，原始文件中各个器官重叠在一起，无法清晰地展示系统和器官，也无法对此进行选择操作。因此研究中就需要对其按照器官或者系统为单位，将图片中的内容在处理后得以分离。分离后所得到的称为原子图片文件。研究中将给出由原始图片加工得到各原子图片的设计步骤详见如下。

(1)从原始 SVG 文件中提取文件头和一些公用的样式作为模板。

(2)遍历 SVG 文件中<svg></svg>标签下的一级子节点(不包括步骤(1)中的样式节点)，调整这些节点中的透明值，全部设成不透明，并将其插入到模板中，得到新生成的 SVG 文件。

(3)将所有新生成的 SVG 文件导出为 PNG 文件。

以上方法所获得的所有原子图片都具有相同的宽度，而在获得所有的原子图片后，便可将每个系统或者器官表示成若干原子图片的叠加。这样，便生成了该系统或者器官所对应需要展示的图片。由此推证得出的整体变换过程如图 2 所示。

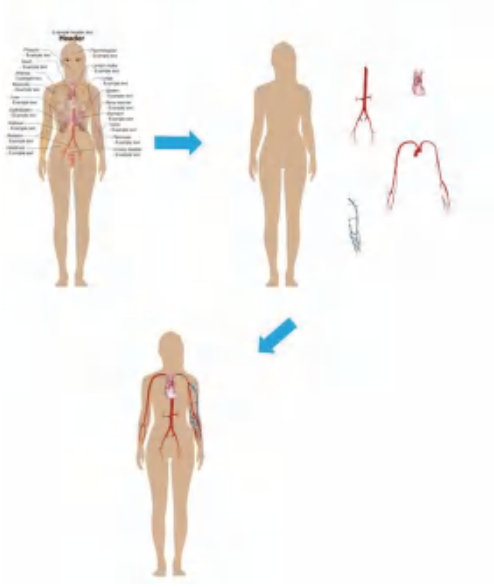

图 2　从原始图片获得心血管系统图片过程示意

Fig. 2　The process of getting the image of cardiovascular system from original file

在获得某系统或者器官的图片后，就可以将疾病本体数据与图片数据之间建立映射关系。基于此，系统就可以在绘制系统或者器官的图片时，也一并罗列出与之相关的疾病本体数据。

2.2　系统或器官点击事件确定

由于人体器官或者系统的图片在绘制时，是一张完整的图片，无法获取其点击的是某一个具体的器官，只能得到鼠标点击的坐标，从而无法对此予以精准响应，以及再对系统或者器官展开进一步的细化，最终对疾病数据实现筛选和过滤。因此，确定鼠标点击位置落在哪个部位的身体器官上是该可视化问题需要解决的核心问题。

本系统通过获取每个器官的边缘路径，判断坐标是否位于这个路径所围成的封闭曲线内部来判断

鼠标是否点击了该器官。若鼠标点击的位置是多个器官的重叠处,则通过事先建立的优先级顺序,决定最终需要响应的器官。

研究时只需确定图片的边缘路径,因此图片内部的具体色彩信息可以被忽略,在这里仅需考虑像素的“有”和“无”两种状态,故而,一个二维图片可以抽象为一个 0-1 矩阵。矩阵中值为 0 处表示该处图片为透明,值为 1 处表示该处存在像素点。综上研究过程可称为图片的二值化。

接下来,拟将引入一些数学定义。对此可研究分述如下。

定义 1 设有 0-1 矩阵 $\boldsymbol{A}_{m\times n}$,当 $2\leqslant i\leqslant m-1$, $2\leqslant j\leqslant n-1$ 时,若同时满足 $\boldsymbol{A}_{i\pm1,j\pm1}=1$, $\boldsymbol{A}_{i,j+1}=1$, $\boldsymbol{A}_{i+1,j}=1$,则称 (i,j) 为 $\boldsymbol{A}_{m\times n}$ 中的内部点。

定义 2 设有 0-1 矩阵 $\boldsymbol{A}_{m\times n}$,若 $\boldsymbol{A}_{i,j}=1$ 且 (i,j) 不是 $\boldsymbol{A}_{m\times n}$ 中的内部点,则称 (i,j) 为 $\boldsymbol{A}_{m\times n}$ 中的边界点。

定义 3 设有 0-1 矩阵 $\boldsymbol{A}_{m\times n}$,若 (i,j) 既不是 $\boldsymbol{A}_{m\times n}$ 中的内部点,又不是 $\boldsymbol{A}_{m\times n}$ 中的边界点,则称 (i,j) 为 $\boldsymbol{A}_{m\times n}$ 中的外部点。

通过对本系统中所使用的图片进行分析,可以发现,所有原子图片中的物体的各个部分是连通的,即所有的内部点是连通的。而且,最多只在物体内部存在若干孔洞,或者分散在整张图片的若干孤立的噪声点。关于噪声点的去除,可做剖析叙述如下。

考虑以某点为中心的 3 * 3 子矩阵内与其相邻的 8 个点,若有大于 6 个点的值与该点不相同,则认为该点是一个噪声点,将其修改为与周围大多数点相同的值即可。而图片内部的孔洞,只需将点击该部分也视作对整个物体的点击,则可以不用对其进行特殊处理。

针对上述情况,还需定义矩阵的边缘路径,内容描述见如下。

定义 4 设有 0-1 矩阵 $\boldsymbol{A}_{m\times n}$,若其所有内部点和边缘点构成的区域是连通的,则称 $S=<(i_1,j_1),(i_2,j_2),(i_3,j_3),\dots,(i_k,j_k),\dots,(i_s,j_s)>$ 为 $\boldsymbol{A}_{m\times n}$ 的边缘路径,其中 (i_k,j_k) 是 $\boldsymbol{A}_{m\times n}$ 的边缘点,且所有的 $\boldsymbol{A}_{m\times n}$ 所有的内部点都在边缘路径 S 所构成的封闭图形内。

简单地说,边缘路径就是图片中物体的轮廓所构成的封闭图形,这是物体边缘点的某种连线方式,该连线方式要求不能穿过物体的内部。为了获取图片内物体的边缘路径,本次系统研发得到的算法流程步骤可详述如下。

(1)从左至右,从上到下遍历矩阵的元素,直到找到第一个点 $P_1(i_1,j_1)$,使得 $\boldsymbol{A}_{i_1,j_1}=1$。并记 $P_0(i_0,j_0)=(i_1-1,j_1)$,同时将 P_0, P_1 加入数组。

(2)取出数组的最后 2 个元素分别为 P_{-2}, P_{-1},计算其向量 $\overrightarrow{P_{-2}P_{-1}}$。

(3)依次遍历点 p_{-1} 的 4 个方向的点,找到第一个边缘点,将其加入数组。遍历的顺序为:以 $\overrightarrow{P_{-2}P_{-1}}$ 为正下方,按照左、上、右、下的顺序完成遍历。

(4)重复过程(2)~(4),直至数组的最后一个元素与 P_1 相同。

(5)设此时数组中为 $<P_0,P_1,P_2,\dots,P_s,P_1>$,则 $<P_1,P_2,\dots,P_s>$ 便是矩阵 $\boldsymbol{A}_{m\times n}$ 的边缘路径。

在此基础上,针对一幅原子图片计算其边缘路径的过程展现则如图 3 所示。

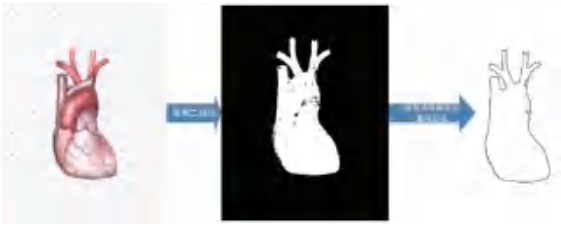

图 3 对心脏计算边缘路径过程示意

Fig. 3 The process of calculating the path of heart image

在计算出边缘路径之后,便可判断鼠标点击时的坐标是否落在边缘路径所构成的封闭图形内部来确认是否点击了该物体。物体点击时的设计效果如图 4 所示。

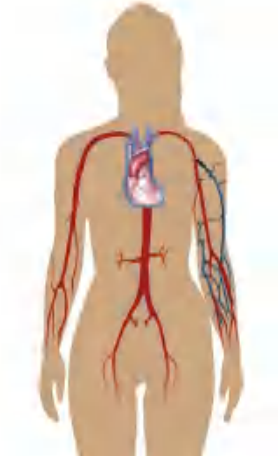

图 4 鼠标点击心脏效果图

Fig. 4 The result of clicking heart organ

3 结果展示

通过将疾病本体数据与图片相映射,以及图片

中相关器官的点击事件的响应，可以构建出一个通过人体生理位置进行相关疾病的检索与筛选的过程。初始状态时，可视化区域展示较高层级的系统或器官，以及与该层级系统或器官相对应的疾病。图 5 即表示了处于高层级状态时的可视化区域内容。而后，通过点击心脏或者血管，可视化区域进入到如图 6 所示的低层级系统或者器官，并对疾病数据进行了过滤。

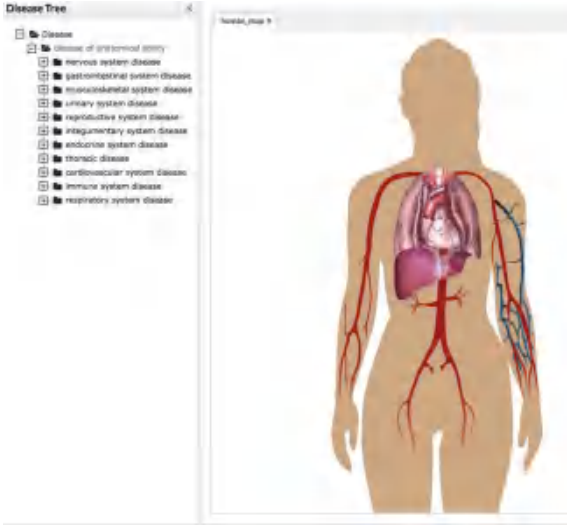

图 5　高层级系统或者器官展示

Fig. 5　The display of high level human system and organs

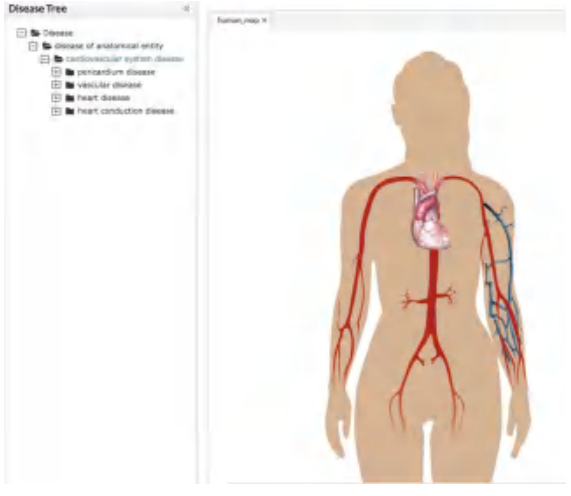

图 6　低层级系统或者器官展示

Fig. 6　The display of low level human system and organs

4　结束语

本文针对疾病本体下的“disease of anatomical entity”分支，提出一种有效的可视化方法，将疾病本体数据映射到人体的生理位置上，并通过在相关系统或者器官上的操作，对疾病本体进行筛选过滤，从而对其做出清晰的表示。

参考文献

[1] SCHRIML L M, ARZE C, NADENDLA S, et al. Disease ontology: A backbone for disease semantic integration [J]. Nucleic Acids Research, 2012,40:D940-946.

[2] LIN Yu, XIANG Zhuoshuang, HE Yongqun, et al. Brucellosis Ontology (IDOBRU) as an extension of the infectious disease ontology[J]. Journal of Biomedical Semantics, 2011, 2(1): 9.

[3] SMITH B, ASHBURNER M, ROSSE C, et al. The OBO foundry: Coordinated evolution of ontologies to support biomedical data integration[J]. Nature Biotechnology, 2007, 25(11): 1251-1255.

[4] TIRMIZI S H, AITKEN S, MOREIRA D A, et al. Mapping between the OBO and OWL ontology languages[J]. Journal of Biomedical Semantics, 2011, 2(1): 1-16.

[5] DEAN M, CONNOLLY D, HARMELEN F V, et al. OWL Web Ontology language 1.0 reference[J]. Anaesthesia, 2004, 59(7): 729.

[6] WANG X H, ZHANG D Q, GU T, et al. Ontology based context modeling and reasoning using OWL[C]// Proceedings of the second IEEE Annual Conference on Pervasive Computing and Communications Workshops.Orlando, FL, USA:IEEE , 2004: 18-22.

[7] PENG Zhongren, ZHANG Chuanrong. The roles of geography markup language (GML), scalable vector graphics (SVG), and Web feature service (WFS) specifications in the development of Internet geographic information systems (GIS)[J]. Journal of Geographical Systems, 2004, 6(2): 95-116.

[8] BATTIATO S, BARBERA G, BLASI G D, et al. Advanced SVG triangulation/polygonalization of digital images[C]//Proceedings of SPIE, Internet Imaging VI. San Jose, California, USA: spie, 2005, 5670(1): 1-11.

[9] GUO Zhimao, ZHOU Shuigeng, XU Zhengchuan, et al. G2ST: A novel method to transform GML to SVG[C]// Proceedings of the Eleventh ACM International Symposium on Advances in Geographic Information Systems. New Orleans, Louisiana, USA: ACM, 2003: 161-168.

[10] BRAY T, PAOLI J M, SPERBERG-MCQUEEEN C M, et al. Extensible Markup Language (XML)[J]. World Wide Web Consortium Recommendation,1998,2(10):35-42 .