**2.新冠概念图谱**

**2.1简介以及潜在应用**

新冠概念图谱从网络文本中采集了与新冠疾病相关的实体和关系（包括疾病、人物、症状等），进一步融合了从百度百科，维基百科等知识库的infobox中提取的实体关系。更为特别的是，本概念图谱利用自动挖掘的手段从网络文本中采集了大量细粒度的上位概念词。即此概念图谱的is-a层次结构（schema）是自动构建而成的，包含了细粒度的上下位层次结构。该概念图谱目前已应用于深睿医疗开发的新冠肺炎小睿医生助手中用于计算问句之间的相似度以及辅助解答用户提问。

借助于细粒度的上位概念层次结构和丰富的实体间关系，本概念图谱能够为针对新冠肺炎的智能服务系统，例如问句相似度计算、智能问答等技术提供数据支持。以问句相似度计算问例，本概念图谱可以对问句中的实体利用上位概念词进行扩展，通过丰富额外的特征从而更加准确的衡量问句间相似度而将问题归类。

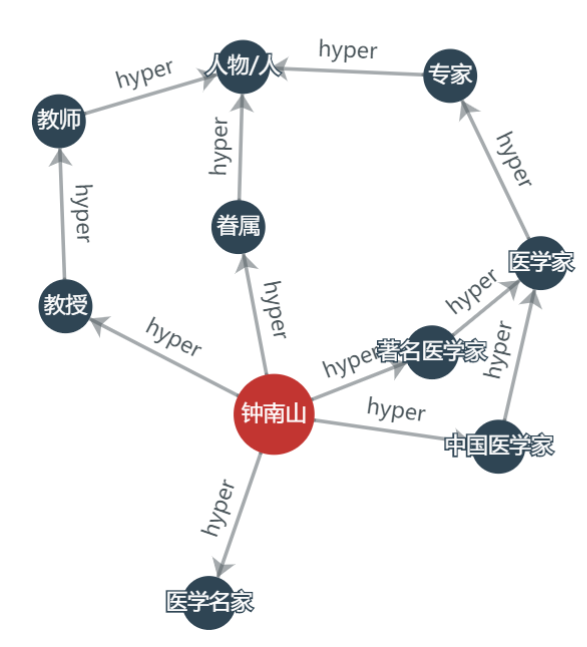
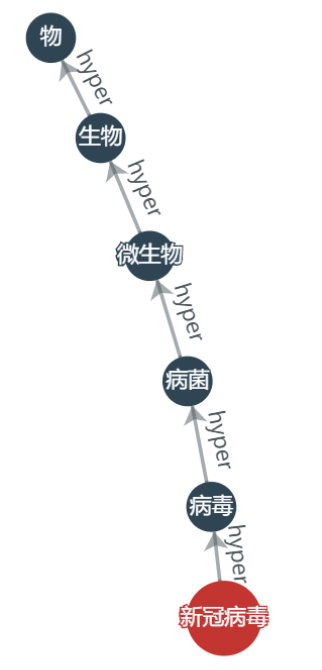


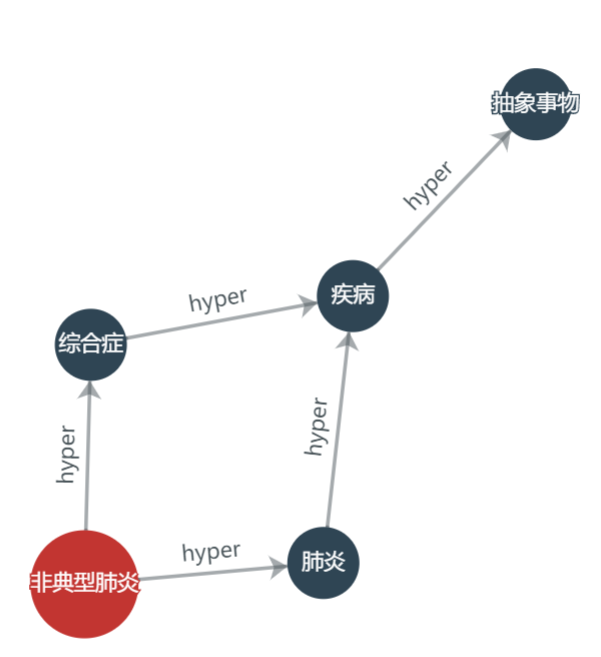
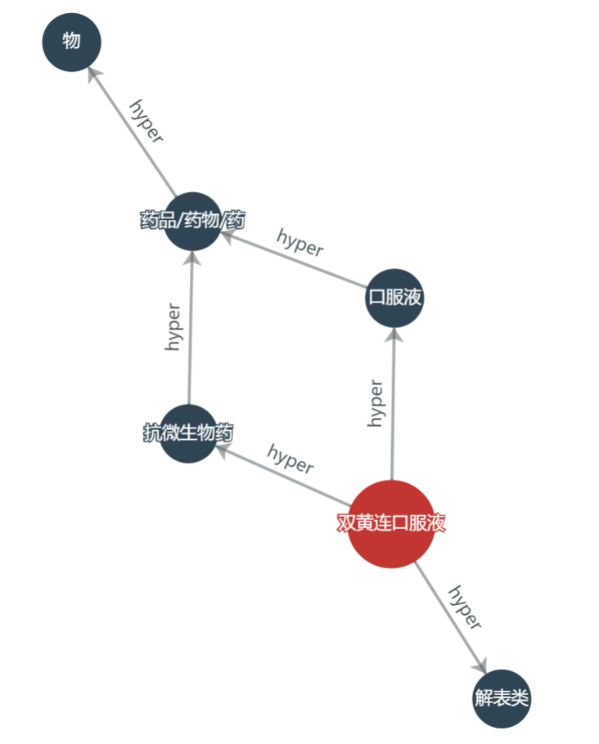


**2.2数据schema**

新冠概念图谱由网络文本中挖掘了大量细粒度的上位概念词，并自动获取概念之间的层次结构。即，本概念图谱的层次结构（schema）是自动构建形成的，可根据实体的含义形成细粒度的层次结构。

下面即展示了针对不同实体所形成的概念层次结构，其中红色结点为实体，深色的为上位概念词。由图中示例可见，根据实体的含义可以自适应的形成不同结构、不同层次的概念层次结构。例：

**2.3数据来源与规模**

新冠概念图谱的数据有两个来源：

1. 从网络文本中采集了与新冠疾病相关的实体、关系、细粒度的上位概念词，自动构建形成上下位概念层次结构。
2. 融合了从百度百科，维基百科等知识库的infobox中提取的实体关系。

新冠概念图谱中的上位概念词从多来源的数据（包括搜索引擎、百科、构词结构）中获得，同时通过排序机制过滤掉可信度较低的上位概念词。实体和上位概念词以及上位概念词之间的层次结构（is-a关系）则通过度量词语之间的语义宽泛性自动获得。横向关系三元组（实体与实体之间的关系）中80%的数据来源于百度百科、维基百科的整合，余下的20%是从网络文本中自动挖掘而得，再辅以句法、语义等规则形成的置信度进行过滤而得到。以上做法保证了新冠概念图谱中包含的上位概念词、概念层次结构和关系三元组的准确性。

**新冠概念图谱规模：**

实体：4784

上位概念词：1487

上下位（Is-a）关系元组：19391

横向实体间关系三元组：35172

**2.5作者介绍**

张裕舟 哈尔滨工业大学社会计算与信息检索研究中心 yzzhang@ir.hit.edu.cn

佘琪星 哈尔滨工业大学社会计算与信息检索研究中心 qxshe@ir.hit.edu.cn

王必聪 哈尔滨工业大学社会计算与信息检索研究中心 bcwang@ir.hit.edu.cn

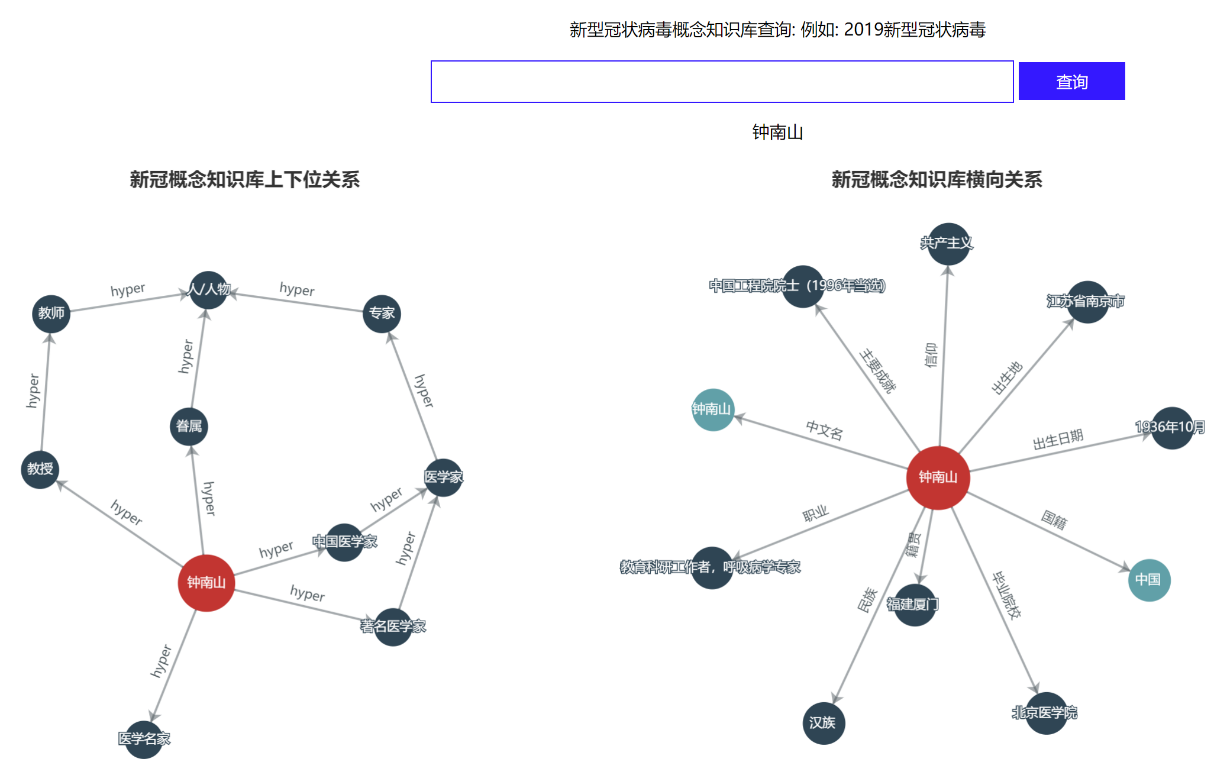
刘铭 哈尔滨工业大学社会计算与信息检索研究中心 mliu@ir.hit.edu.cn

秦兵 哈尔滨工业大学社会计算与信息检索研究中心 bqin@ir.hit.edu.cn

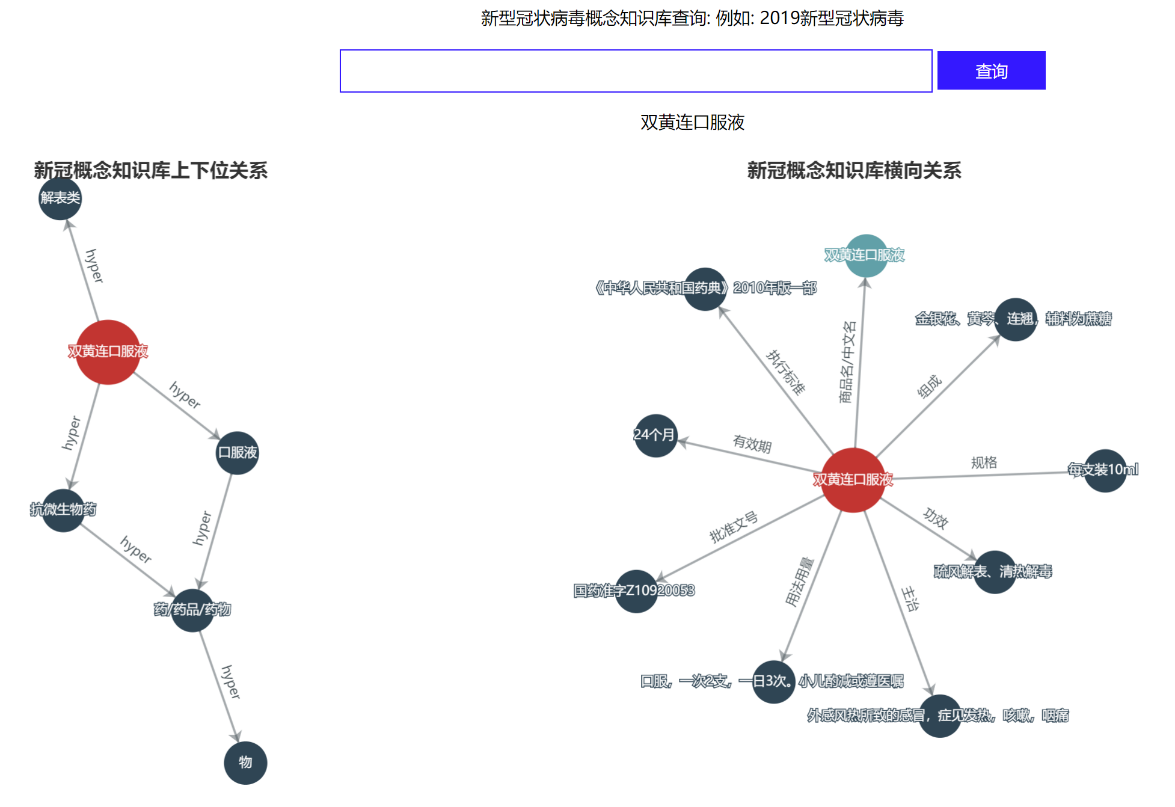
刘挺 哈尔滨工业大学社会计算与信息检索研究中心 [tlu72@ir.hit.edu.cn](mailto:tlu72@ir.hit.edu.cn)

**2.6可视化样例**

新冠概念图谱除了开放数据之外，还提供了可视化的查询网站，通过输入的关键词能够在该图谱中查询实体。为方便展示实体的上下位概念层次结构和实体之间的横向关系，我们提供了两种展示样式。如下图所示，针对同一实体的两种展示样式分别是实体的上下位概念层次关系和实体间的横向关系。图中红色的点代表查询实体，查询实体周围浅绿色的点代表该结点对应的实体也被概念图谱所包含，可点击展示。



same entity



same entity