·业务研究·

基于主题图的在线健康信息标签语义挖掘研究

占　泓,熊回香,蒋武轩,李　琰

(华中师范大学 信息管理学院,湖北 武汉 430079)

摘　要:【目的/意义】在线健康信息的有效组织对提升全民身体素质具有重要的社会价值。【方法/过程】在分析健康信息主题、关联关系和资源标引的基础上,构建基于主题图的在线健康信息标签语义挖掘模型,从而构建了健康信息标签主题图并实现了其可视化导航、浏览和检索等功能。【结果/结论】基于主题图的在线健康信息标签语义挖掘模型能够准确的发现在线健康信息与信息标签间的深层关系,可以更好地揭示在线健康信息标签的语义关联,为用户提供信息的可视化浏览和导航功能、提升健康信息的组织效果,帮助用户健康信息获取。【创新/局限】本文将主题图与健康信息标签相结合,提高了健康信息的检索效率和利用效率,但本文也存在着不足,例如标签样本量和样本范围较小,缺乏专业医学研究者的参与。

关键词:健康信息;标签;语义挖掘;主题图;Ontopia

中图分类号:G254　　**DOI:**10.13833/j.issn.1007-7634.2022.01.017

1 引　言

随着社会经济和技术的不断发展,人们对生活的追求从满足基本生存需要转变成提升生活品质、注重身体健康的方向发展。为了满足人们对健康的需求,大批健康网站应运而生,与之而来的是健康信息急剧增长,用户对健康信息的有序组织提出更高的要求。社会化标注凭借其在网络信息组织中的优势被大范围的应用在健康网站中,通过用户自发地定义与健康信息有关的标签来对健康信息进行标注组织,不仅进一步优化健康信息的组织,也更加符合用户的认知习惯,满足用户健康信息查询的需要,提高了健康信息的利用率。因此,我们把在线健康社区中专门标注健康信息的标签称为健康信息标签,这类标签能够一定程度对健康信息组织提供基础。但随着健康信息及健康信息标签的不断增长,标签的共有问题在健康信息标签中逐渐显现,如标签歧义、模糊、冗余现象,这些问题降低了健康信息标引和检索的有效性[1]。而主题图比标签更为规范、准确,并且更为直观,利用主题图可以改善健康信息标签存在的问题。

主题图在语义挖掘领域常被用于知识的组织,其对关联关系的挖掘和对主题的层级划分能够更加准确的发现信息的语义信息和等级关系,较好的解决了健康信息间的层级结构模糊和健康信息标签的语义缺失等问题。因此,针对健康信息层级关系难以发现和健康信息标签语义缺失、难以与专业医学词汇匹配等问题,本文创新性的将用于知识表示的主题图引入健康信息组织研究中,融合标签和主题图技术构建基于主题图的健康信息标签语义挖掘模型,以期改善健康网站的语义挖掘、集成、检索和导航的效果,规范健康信息标签质量,提高用户组织、检索信息的效率,促进健康网站信息系统的良性循环。

2 相关概念及研究现状

2.1 健康信息标签及其研究现状

(1)健康信息标签

标签被广泛应用于网络信息组织中,通过对网络信息的准确标注,将繁杂的网络信息予以分类整理,保证了网络信息的组织和用户的信息查询效率。而在线健康社区作为一类具有特殊健康信息的网络载体,对其信息的标注尤为重要。因此,我们把在线健康社区中专门标注健康信息的标签称为健康信息标签。目前基于健康信息标签的分类法已经被作为一种有效的信息组织方式被运用于健康网站中。

健康信息标签对健康信息的有效表征能更好地反映在

收稿日期:2020-11-25

基金项目:华中师范大学中央高校基本科研业务费(人文社科类)重大项目"基于语义网的在线健康信息的挖掘与推荐研究"(CCNU19Z02004)。

作者简介:占泓(1995-),女,湖北黄冈人,硕士研究生,主要从事网络信息组织与检索研究;熊回香(1966-),女,湖北鄂州人,博士,教授,博士生导师,主要从事网络信息组织与检索研究;蒋武轩(1993-),男,吉林白城人,博士研究生,助理会计师,主要从事网络信息组织与检索、信息化研究,通讯作者;李琰(1995-),女,湖北武汉人,硕士研究生,主要从事电子商务研究。

线健康信息的主要内容，同时对健康用户需求的表征也更加准确。在在线健康社区中，健康用户对在线健康信息能随时赋予全新的标签，根据追踪健康信息标签的变化，可以把握用户关注健康问题的相关进展，更能满足健康用户对健康信息的需求。

（2）相关研究

当前国内外针对健康信息领域标签的研究较少。在科研方面，国外针对健康信息的研究，较早主要集中在利用语义网、本体等技术来发现健康信息间的语义关系，从而对健康信息进行有效组织。如生物医学数据语义集成平台Linked LifeData利用有效的推理在一定程度上缓解数据之间的语义问题，并最终用类似的数据模型进行聚合，形成异质生物医学数据[2]；Noor等学者利用语义网技术来研究文献中药物之间的关系[3]；Huang等人提出并构建了基于本体的领域特异性知识库来提高RNA目标基因预测的知识获取效率[4]；Tenenbaum等人提出基于本体技术，将生物医学相关数据建立成本体结构，从而挖掘其中的隐性语义关系，构建生物医学知识发现系统，为研究人员同时进行临床研究和基础研究提供技术支持[5]；Kardan等学者则是利用WordNet来建立相关标签的语义层级，有利于消除标签之间的同义问题[6]；而国内学习借鉴国外相关研究经验，也多从本体和语义网方面进行研究。如张军亮将中文医学术与UMLS(医学一体化语言系统)进行映射，对其语义概念进行关联的基础上设计语义解释空间，从而构建基于语义关联的多源医学信息资源发现服务系统模型[7]；张传文基于本体(Ontology)和语义网(Semant-ic Web)的融合，采用将本体混合的方法构建语义模型，利用Hadoop分布式数据处理架构来进行数据处理，通过整个服务体系来构建全局本体，对各个数据源与相对应的本体进行定义，使语义模型具有更好的可拓展性和统一性，从而提出医疗信息采集和共享的新架构[8]；但杨帅旗创新的以使用标签整合相关医疗资源的健康网站PatientsLikeMe为研究对象，提取网站中的主题词集与健康标签，然后根据两者间的相似程度来划分层级，最终形成富有语义内容的树状主题标签集，基于公众分类的医疗门户网站分类系统，对原有的分类体系进行系统性改进[9]。

在健康信息标签实践方面，国外最早利用标签对医疗相关资源整合的是PatientsLikeMe和TuDiabetes.org网站，PatientsLikeMe是一个给用户提供在线交流功能的社区，用户可以借此了解最新的医疗信息和进行医疗信息分享，通过用户对自身情况的标签标注，有助于专业医务人员的临床分享。而TuDiabetes.org通过对用户注册时关注的标签向用户推荐相关医疗信息，并促进其参与交流。国内目前较为知名的网站是39健康网进行健康信息标签的采集，39健康网在旗下子网站就医助手、药品通、名医在线、39问医生、疾病百科中都利用标签进行信息标注并提供给用户导航。

综述所述，国内外极少有对在线健康信息及健康信息标签的研究，仅有的研究主要是利用相关技术对医疗信息两者的语义层面进行探究，均未从不同语义间的层级关系进行挖掘以及在更深层的知识表示层面来探究两者间的关系。

2.2 主题图技术及其研究现状

（1）主题图技术

在ISO标准中(ISO/IEC13250)介绍主题图为：通过定义一个分布许多代表主题结点的空间，利用空间中不同结点间连接需要经过中间结点的数量来衡量不同主题间的距离及描述不同主题间建立关系的路径[10]。主题图由三个基本要素组成：主题、关联、资源[11]。其中，“主题”是将一切客观存在进行抽象后生成的代表性内容，主题是按照类型对其进行划分，“关联”代表两个或多个主题之间的关系，这种关系可以具有多种类型并同时存在；而“资源”即主题连接的与之相关的一种或多种信息资源。

主题图的概念模型包括信息资源层和主题地图层两层。信息资源层为互联网中各类信息资源的集合，而主题地图层是基于信息资源层而构建起来的主题网络，以资源指引相连接，由主题和主题间关联关系所组成，如图1所示。主题地图层中的黑色结点代表着不同主题，主题间的连线代表着不同主题间的关联关系，而主题地图层通过资源指引到达信息资源层，信息资源层包含着不同的信息资源，包括网页、文章、视频等等，这些信息资源与主题地图层的主题相联系，又相互独立的存在互联网中[12]。

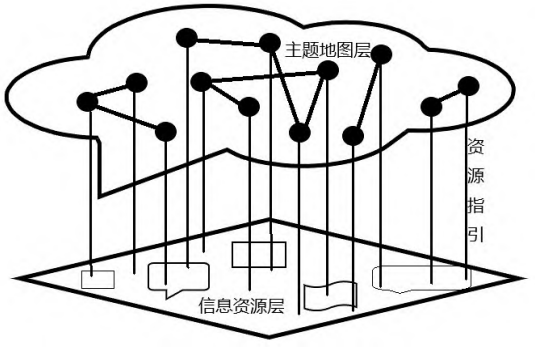

图1 主题图概念模型

Figure 1 Topic map conceptual model

目前，大多主题图应用开发工具都基于TMAPI或主题图引擎来进行开发，为其开发人员提供生成、更新、储存、展示服务。其中最经典的为支持TMAPI的Ontopia知识开发组件(Ontopia Knowledge Suite，OKS)，OKS是Ontopia公司开发的并提供给主题图创建人员使用的相关资源组件[13]。OKS是目前能用于创建、维护、配置主题图的最完整并且功能最齐全的开源主题图创建工具。

（2）研究现状

当前国外学者大多利用主题图来优化资源间的语义结构。如Rath认为利用主题图技术可以让主题、关联关系和资源标引构成主题图模型，同时说明关联这一因素对主题图中语义关系的揭示的重要性[14]；Wang等提出了基于主题图的语义知识融合框架，通过计算知识对象与主题的语法相似度和结构相似度，将两者的相似度与权重结合起来得到整体

表1 健康信息标签展示(部分)

Table 1 Health information label presentation (Part)

咨询文本	患者病情概述	健康信息标签
1	眼睛不舒服	高血脂、糖尿病、高血压、血管、…、眼睛
2	睡觉大便失禁	糖尿病、大便失禁、入睡、肠鸣、…、胃
3	右脚踝关节肿痛	糖尿病、痛风性关节炎、关节炎、踝关节扭伤、…、痛风
4	判断糖尿病类型	糖尿病、一型糖尿病、二型糖尿病、血检、…、检查结果
5	干细胞移植有用吗	糖尿病、尿毒症、脑梗、消极、…、无力
6	血糖情况询问	糖尿病、胰岛素、糖化血红蛋白、血糖、…、保健品
7	伤口难愈合	伤口、愈合难、糖尿病、感染、…、体质
8	糖尿病肾衰竭	电解质紊乱、下肢浮肿、肾病、肾功能衰竭、…、肾衰
9	是否用药	糖尿病、米诺地尔曼迪、斯必申、病人、…、用药
10	免疫性糖尿病患者	糖尿病、低血糖、胰岛素、血糖、…、黄瓜
…	…	…
2050	糖尿病脑血栓	高血压、糖尿病、脑血栓、桦树茸、…、并发症

语义相似度,一定程度解决了知识融合结构异质性和语义异质性问题,完成语义层面的知识融合[15];Charles Robert等提出将主题图技术运用于注释信息,用对象、资源指引、地点、标注者、描述、时间等维度来建立注释与资源、注释与主题间的语义关系,从而建立有语义关联的知识数据库,从而加强检索结果的语义相关性[16]。在国内,王平、吴玉萍认为基于主题图来进行数字图书馆的知识组织活动是可行的,同时构建了数字图书馆知识组织模型[17];李纲、王忠义提出了将主题图应用于隐性知识的管理方面,扩充了隐性知识相关的理论[18];夏立新、徐晨琛设计了以主题图技术为基础应用于政务门户信息的导航系统[19]。

综上所述,国内外已有一些关于健康信息和标签的语义挖掘研究,但还有些不足:第一,健康网站中关于健康信息标签的利用较少,国内外对于健康信息标签的语义挖掘研究也较为不足,健康信息标签仍存在着语意不明、扁平化、与专业医学词汇难以匹配等语义问题。然而健康信息标签语义层次的挖掘对健康网站建立符合健康信息用户日常语言习惯的检索和导航起着重要作用,因此基于主题图对健康信息标签进行语义挖掘是很有意义的;第二,与其他语义挖掘方法相比较,将主题图应用于健康信息和标签的语义挖掘研究不多,而主题图能在一定程度上弥补标签的语义问题,所以有关这方面的研究亟待扩展深入。

3 数据来源与模型构建

3.1 数据来源

本文选取了39健康网进行健康信息标签的采集,39健康网在旗下子网站就医助手、药品通、名医在线、39问医生、疾病百科中都利用标签进行信息标注并提供给用户导航,2017年39健康网获得了知识产权管理体系认证证书,其网站的内容质量较好,也为标签的质量提供了保障。39健康网对糖尿病专题也进行了主题分类,因此本文依据网站已有的分类对一、二级主题类型进行定义。

本文以39健康网中39问医生板块糖尿病专题中的标签为实例,首先对健康信息标签资源进行采集清洗,根据39健康网对糖尿病专题既定的主题分类来对标签资源进行分析并匹配到相应的主题类型中。再用Ontopia环境中的Ontology工具进行主题的生成、关联关系及名称类型的定义,其后利用Ontopoly中Instance工具将标签输入,并用相应的关联关系指定标签间、标签与主题间的关系,生成标签主题。用Ontopia环境中Omnigator组件来实现健康信息主题的浏览、导航、更新功能。最后通过主题图语义检测和主题合并功能实现健康信息标签主题图的最终建立,利用Visual板块实现主题图的可视化,实现健康信息标签主图语义关系的显性化。

相关资源是通过集搜客相关爬虫工具进行爬取,所爬取的文本包括患者咨询文本和医生回答文本两部分。为了保证所提取的健康信息标签能完整展示和表达健康用户提问和医生回答全过程,本文除了对咨询文本的标签进行爬取,还利用集搜客中的分词软件对未提供标签资源的医生回复文本进行分词,自动提取其中的健康词汇,接着人工对这些健康词汇进行清洗和筛选,并手动添加相关词汇来完善词汇集,将这些健康词汇作为对健康信息标签集的补充,再手动对所有的健康信息标签进行去重、去除停用词等预处理,最终获得2079个健康信息标签。由于所收集到的健康信息标签集数量较大,本文只对部分健康用户的提问和健康信息标签进行展示,见表1。

3.2 模型构建

本文的健康信息标签主题图体系结构在主题图技术的支持下,将健康信息标签融合于主题图结构中构成。分别从健康信息标签资源层、健康信息标签主题图生成层、应用层以及用户层四个层次构成了基于主题图的在线健康信息标签语义挖掘模型,如图2所示。

(1)健康信息标签资源层

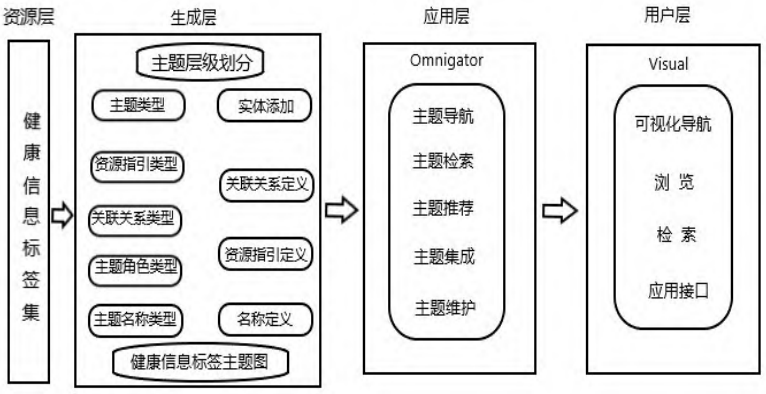

图2 基于主题图的在线健康信息标签语义挖掘模型

Figure 2 Topic map based semantic mining model of online health information tag

表2 糖尿病专题主题类型展示(部分)

Table 2 Presentation of thematic types of diabetes (Part)

一级主题类型	二级主题类型	健康信息标签实例
糖尿病类型	1型糖尿病	儿童、I型、小儿糖尿病、尿床、…、遗传
	2型糖尿病	Ⅱ型糖尿病、老人、肥胖症、高血压、…、高血脂
	妊娠期糖尿病	怀孕、巨大儿、羊水过多、流产、…、难产
	继发性糖尿病	胰腺炎、分泌疾病、胰脏疾病、胰岛细胞、…、基因缺陷
糖尿病症状	身体表现	三多一少、体重下降、出汗多、口干、…、多尿
	血糖指标	产后血糖、异常血糖值、空腹血糖、糖化血红蛋白、…、餐后2小时血糖
	尿糖指标	尿比重、尿糖阳性、血尿素氮、尿酮体阳性、…、尿酸高
	并发症	糖尿病足、肾病、心血管疾病、胃肠疾病、…、脾虚
糖尿病治疗	饮食疗法	谷物、水果、粗粮、茶水、…、青菜
	运动疗法	游泳、散步、慢跑、骑单车、…、瑜伽
	胰岛素注射	上臂注射、大腿注射、来得时厂家、臀部注射、…、胰岛素注射液
	口服药物治疗	三肾丸、卡托普利片、格列美脲滴丸、格列齐特片、…、盐酸二甲双胍缓释片
	手术疗法	肾移植、胃束带、干细胞移植、外科手术、…、人流
糖尿病护理	禁忌	可乐、蜂蜜、淀粉、油腻、…、高盐
	生活起居	拔牙、热水泡脚、针灸、滑石粉、…、睡眠
	每日自测	DQ位点多态性关系、c肽释放试验、产后体内胰岛素拮抗激素水平、尿蛋白定量、…、血糖监测
	心理	焦虑、失眠、多梦、坐立不安、…、抑郁
	家属	知识普及、饮食、心理疏导、支持、…、家族史

健康信息标签资源层是基础层,是构建起健康信息标签主题图的第一步。该层存储了用于生成主题的所有健康信息标签,这些标签都来源于39健康网的39问医生板块糖尿病专题的2502条医患问答文档中,为了使健康信息标签对整体糖尿病领域有着系统的表达,本文的标签集不仅包括了健康用户标注的标签,还从医生回复中经过分词、筛选,提取了500条健康信息标签,经过去重、去除无效标签、整理后最终形成了包含2072条健康信息标签的标签集。

(2)健康信息标签主题图生成层

本层是健康信息标签体系的核心层,将分别对健康信息标签主题图的主题类型进行划分,并进行关联类型和资源指引。

①健康信息标签主题图的主题类型划分

本文对一级主题和二级主题的定义参照39健康网对糖尿病已有既定的主题划分来进行,这样更符合健康网站对于糖尿病领域主题的分类,也与糖尿病在线资源的实际使用更为贴近,使得健康信息标签主题图更符合健康用户的使用习惯。本文为"糖尿病"健康信息标签主题图对象先定义了四个一级主题类型,其次,在一级主题类型下再定义了18个二级主题类型,如表2所示。

②健康信息标签主题图的关联类型

将健康信息主题确定后,要通过关联关系的定义来挖掘主题之间的内在语义关系,因此要定义主题间的关联关系的类型。本文选取了"糖尿病"标签主题图中的"尿糖阳性"这一主题,来展示"尿糖阳性"主题与其他相关联主题的关系,其中包括了属于关系、包含关系、近义关系、辅助关系、治疗

关系、具体表现关系、禁忌关系,如图3所示。

③健康信息标签主题图的资源指引

由于本文是基于39健康网的问答板块中的内容建立起来的健康信息标签主题图,因而在进行资源指引时,将每一个实例主题都链接到了与其相关的问答内容上,这些内容都是通过网址链接进行连接,查询时会跳转到相关网页中,问答板块的内容多以文本和图片为主,因此本文在对资源指引类型进行定义时,定义了"超链接"这一类型,表示这一资源指引时用来跳转到网页中的。"图片"这一资源指引类型说明该链接所连接到的是图片资源。"文本"这一资源指引类型,表示该健康信息标签主题链接的是基于文本的医患问答。

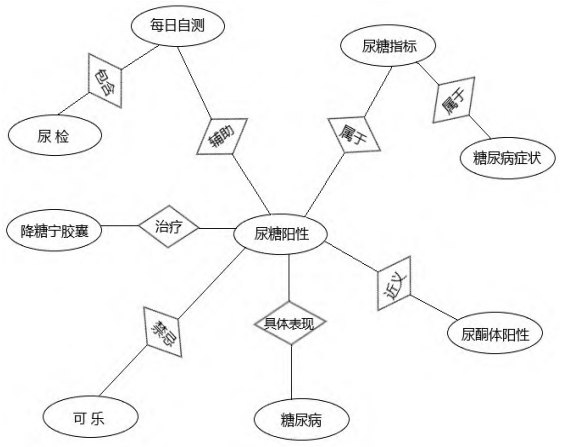

图3 "尿糖阳性"主题关联类型图

Figure 3 Topic association type map of "Positive urine sugar"

(3)健康信息标签主题图的应用层

应用层由五个模块组成,分别是主题导航模块、主题检索模块、主题推荐模块、主题集成模块和主题维护模块。

①主题导航模块。主题导航模块中所有主题的一级主题用被表示为根节点Root,Topic Type表示基本主题类型,而在对主题类型进行添加操作时,可以同时指定它的上级主题和下级主题,因此最终形成的健康信息标签主题图中的主题都是按照层级关系进行排列的。

②主题检索模块。在主题检索模块中健康用户可以通过输入检索语句,进行主题的查询,Omnigator中的检索组件就通过主题匹配来定位到相应主题,并通过用户界面回馈给用户。在Omnigator中,检索行为要基于主题图查询语言(TMQL)来实现,该语言在输入想要查询的主题时可以同时限定其主题范围和关联类型,如此一来就能更准确地定位到主题资源,为用户节省检索时间 。

③主题推荐模块。由于健康信息标签主题图中的主题间具有很强的层级关系和语义关联关系,健康用户在进行查询时,这一查询主题就能通过层级关系找到同一层级资源,或者通过用户所制定的关联关系类型来获取跟检索主题有着这一关联关系的其他所有主题,进而对用户进行推荐。

④主题集成模块。主题集成模型是基于标签主题的语义信息来对标签进行划分和集合,从而更好地对标签进行主题生成,这种集成模式能使用户快速检索到与查询标签主题语义相关的健康主题,也有利于更好地提供推荐服务,使用户在获取健康资源的同时理解其内涵和该主题所在的知识结构。

⑤主题维护模块。主题维护模块包括添加主题(Add)、删除主题(Delete)、更新主题(Update)三个基本功能,并且还支持对主题(Topic)、主题类型(Topic Type)和关联关系(Association)的更新。

(4)健康信息标签主题图的用户层

用户层有可视化导航、浏览、检索、提供外部交互接口的功能。在健康信息标签主题图中,用户可以通过Visual模块的VizDesktop插件实现对健康信息标签主题图的动态浏览。主题图以彩色的界面显示,不同颜色的健康信息标签主题代表着不同语义层次,不同的关联关系也用不同的颜色的线条显示。mnigator组件为用户提供健康信息标签主题图的主题页面浏览,通过健康主题索引展示健康信息标签主题图中主题的基本层级结构,通过查询功能(Query)可快速的获取用户所需健康主题及其关联主题。用户明确该主题的主题范围后,可通过相同主题范围进行反馈推荐。

4 健康信息标签主题图实证

4.1 健康信息标签主题图构建

Ontology由本体编辑器(Ontology Editor)、实例编辑器(Instance Editor)两部分组成。本体编辑器提供对主题类型、资源指引类型、关联关系类型、名称类型、角色类型的定义,最终生成富有层级结构的健康主题集,为实例编辑器搭建起层级框架。实例编辑器则对输入的标签进行名称、资源指引、资源链接、标签所属主题类型、标签间的语义关联关系的输入,最终形成完整的健康信息标签主题图语义知识结构。

(1)健康信息标签主题图本体生成

在本体编辑器Otology中,对"糖尿病症状"这一一级主题类型进行定义,分别定义了其名称、主题身份、主题描述、其上下级主题和其包含的关联关系类型。其中资源指引为39问医生与糖尿病症状相关的所有问答的链接;主题描述对"糖尿病症状"进行了大致的介绍说明;其下级主题类型分别为"尿糖指标""并发症""血糖指标"和"身体表现";由于其位于一级,所以没有上级主题类型;在"糖尿病类型"这一一级主题所包含的健康信息标签实例主题中,定义了"近义""包含""被包含""被治疗""病症具体表现"等关联关系,这些健康标签实例主题可以通过这些关联关系来揭示其语义关系。

(2)健康信息标签主题图实例生成

在确定健康信息主题类型后,通过实体编辑器Instance对健康信息标签主题类型添加实例。例如将标签实例主题名称定义为"脑梗塞";其超链接为39问医生中所有与脑梗塞相关的问答信息的网页链接;与在一级主题类型"糖尿病

图4 基于Omnigator的“禁忌”主题类型界面

Figure 4 Omnigator-based "Taboo" topic type page

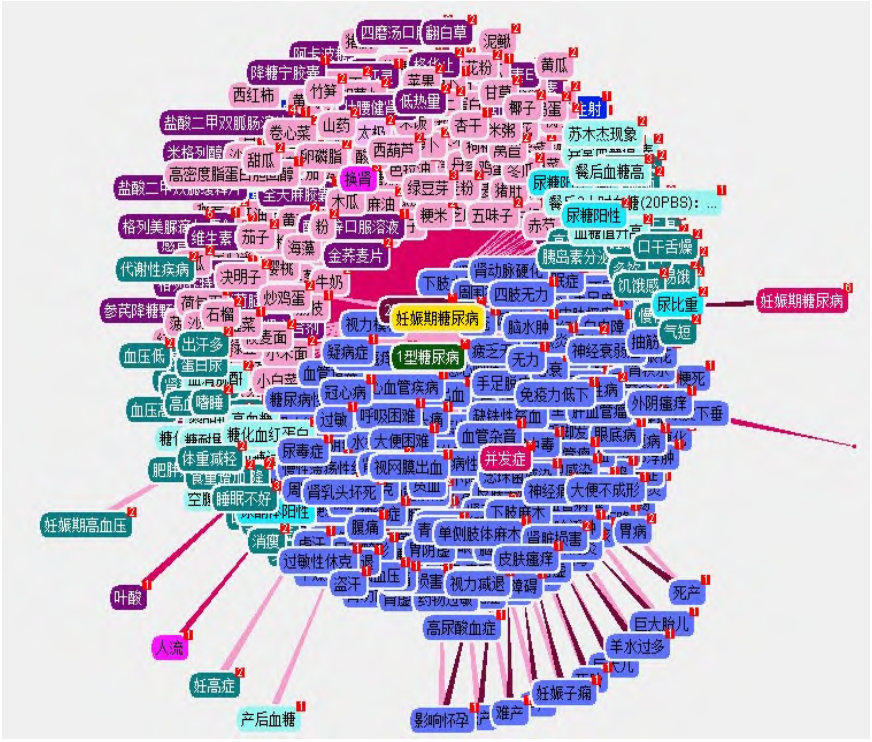

图5 健康信息标签主题图部分可视化展示

Figure 5 A part of the health information label topic map visualization

症状”下的标签实例主题“脑梗死”为“近义”关系，和“脑血管病”存在“被包含”关系；“脑梗死”病症属于“被治疗”范畴；“脑梗死”症状多发于“1型糖尿病”“2型糖尿病”和“继发性糖尿病”中。

4.2 健康信息标签主题图浏览

Omnigator是Ontopia环境中的主题图浏览器，通过Web应用接口，以HTML格式实现主题图的界面展示，用户可在Omnigator主页上直接浏览所有主题的层级和名称、资源指引类型、主题数量等信息，如图4。该浏览界面以网页的方式展示了“糖尿病”中“禁忌”主题类型的关联类型和包含的主题实例。

4.3 健康信息标签主题图的可视化

主题图的可视化是对已经生成的主题图进行一种动态形式的展示，它将一级主题类型作为一个核心节点，通过不

同颜色的连线表示与主题类型连接着的不同类型语义关联关系,而这些线条以发散状分布,另一端连接的节点也用不同颜色表示出来,代表着不同层级的标签主题,这些主题都可以实现扩展节点(Expand node)功能,在原主题图基础上再生成与扩展节点相关的其他标签主题。本文实例中的健康信息标签主题图利用Ontopia Visual Navigator可视化组件实现,采用动态网状图形方式展示健康信息标签中潜在语义关系,为健康用户提供更加直观的浏览、查询界面。在每个健康信息标签主题上面都有其相关标签主题的数字,反映出健康信息标签关联维度,健康用户可以根据需求选择健康信息标签主题进行扩展查询浏览,在遇到多余信息时也可利用隐藏节点功能(Hide node)来折叠该部分健康主题信息,如图5。图中展示了“糖尿病”中“糖尿病类型”“糖尿病治疗”“糖尿病症状”等主题关系的可视化图,健康用户可以通过对健康标题主题的关联实现进一步查询,获取更多语义层面的相关健康主题资源,也可以直接用语句进行主题查询,这种可视化提高了健康信息的查准率,也增强了系统和用户间的交互。

5 健康信息标签主题图语义挖掘结果分析

本文对同一健康信息标签标注的不同主题(主题共现)、同一主题不同健康信息标签(标签的共现)、不同健康信息标签主题类型间的关系对健康信息标签主题图的可视化进行展示,一定程度上解决了健康信息标签的语义问题。

5.1 同一健康信息标签标注的不同主题

对同一健康信息标签标注不同主题的情况进行探究,实例结果如图6所示,图中可以直观地看到“胰岛素注射液”标签链接了这种治疗方式可以治疗的所有糖尿病类型,如“1型糖尿病”“2型糖尿病”等,也展示了辅助治疗的C-肽指标和血糖仪,同时还说明这种治疗方式能对精神有着辅助治疗作用。也展示了“胰岛素注射液”的上级主题类型和与其上级主题同一层级的其他主题类型,可以清楚看出这些健康主题间的语义层级和语义关联。这种网状结构更加直观地展示了健康信息标签主题图的语义信息,这种可以不断扩展的主题图展示形式,也对用户从整体上掌握健康信息标签主题图语义结构有着积极作用。

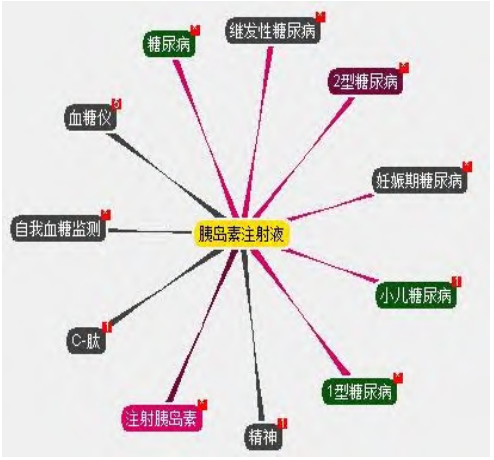

图6 “胰岛素注射液”健康信息标签关联主题可视化展示

Figure 6 Visualized display of related topics on health information label of "insulin injection"

5.2 同一资源的不同健康信息标签

对同一资源的不同健康信息标签进行探究,实例结果如图7所示,图中选取了糖尿病的口服药物“参芪降糖颗粒”这一标签实例。可以看出,“参芪降糖颗粒”可以治疗“继发性糖尿病”“妊娠期糖尿病”和“2型糖尿病”这几种“糖尿病”类型;在其所属的上级主题“口服药物治疗”中,又有近义标签主题“米格列醇片”“四磨汤口服液”“格列美脲滴丸”等;“每日自测”下级主题“自我血糖监测”“血糖仪”“空腹”等辅助了“参芪降糖颗粒”对糖尿病的治疗;“参芪降糖颗粒”最主要的受众是“老年人”。这种健康信息标签主题图可为用户提供直观高效的标签资源检索和导航,并且更形象地展示了健康信息标签间的语义关系。

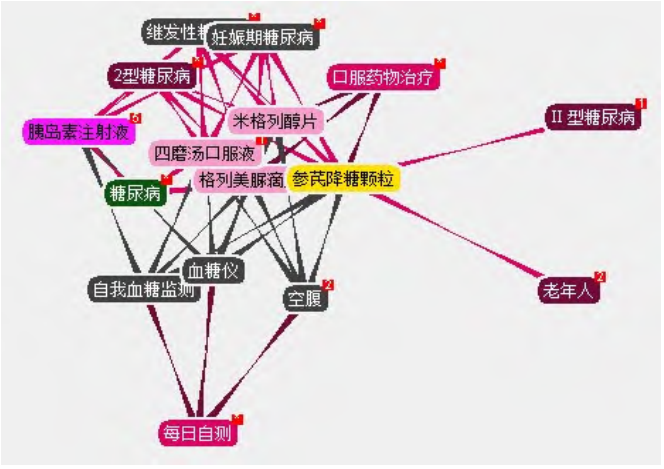

图7 “参芪降糖颗粒”连接标签的可视化展示

Figure 7 Visualized display of the labels connecting with "Shenqi Jiangtang Granules"

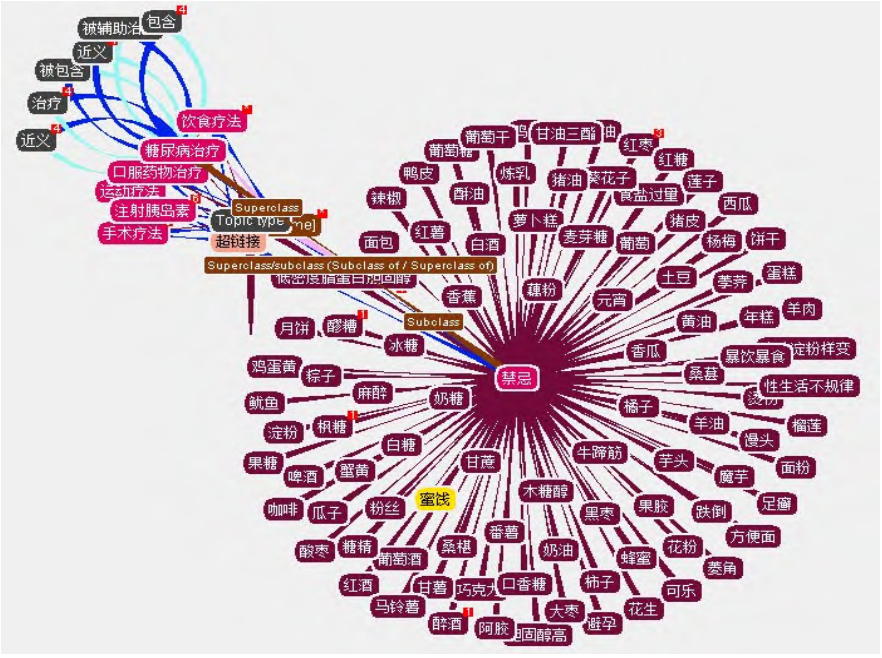

图8 “禁忌”主题所包含的标签主题关联关系可视化展示

Figure 8 A visual display of the tag topic associations contained in the “Taboo” topic

5.3 不同健康信息标签主题类型之间的关系

对不同健康信息标签主题类型之间的关系进行探究,实例结果如图8所示,图中将实例中与“禁忌”有关的健康信息标签主题通过包含关系、被辅助治疗关系、近义关系、被包含关系、治疗关系聚合在一起。

“禁忌”主题包含了“禁忌”主题的“藕粉”“元宵”“香蕉”,体现了“包含”与“被包含”关系;“饮食疗法”“口服药物治疗”“运动疗法”“注射胰岛素”“手术疗法”与“糖尿病症状”主题的下级主题都为“治疗”关系;“冰糖”“白糖”“枫糖”“蜜饯”等都为“近义”关系。

通过这些关联关系的相关维度,对其进行追踪,用户可以直接点击该关联关系,得到与“禁忌”主题通过该关联关系链接的其他主题,得到更加详细的可视化主题图。由此可见将不同健康信息标签主题之间通过不同类型的关联关系进行连接,可以实现对有相关关系主题的组织和聚合,从而形成实例中“禁忌”知识的知识网络。由此可以看出,将主题图运用于健康信息标签可以建立语义层级明确、语义丰富的健康知识体系。

6 结 语

本文将主题图与健康信息标签相结合,构建了基于主题图的在线健康信息标签语义挖掘模型,并构建了健康信息标签主题图及其可视化。一方面基于健康信息标签语义信息的健康信息标签主题图,为健康用户提供了语义导航,这种导航能基于用户的检索主题呈现出其相关主题、同一层级主题和其上下层级主题,一定程度上解决了健康信息标签的层级不明、语义模糊等问题;另一方面健康用户通过主题图工具还能对其检索主题进行可视化浏览,这种浏览方式有利于用户更好理解自己所检索的健康信息,从而提高健康信息的检索效率和利用效率。本文以39健康网中39问医生板块糖尿病专题中的标签为数据来源进行实证,利用主题图工具建立了“糖尿病”健康信息标签主题图,验证了其对于语义挖掘的可行性。但目前的研究还存在不足之处,日后的研究方向应该继续扩大健康信息标签样本量和样本范围,将其它健康资源加入到主题图知识体系中;其次与医学专业学者合作开展研究,使得语义关联更加符合医学领域共识,更精准地挖掘健康信息间的语义关系,使得健康信息标签主题图能够更好地应用于医学领域。

参考文献

1 Arekar T,Sonar M R S,Uke N J.A Survey on Recommendation System[J]. IOSR Journal of Comp-uter Engineering,2015,5(1):1-4.

2 Ontotext.linked life data[EB/OL].[2020-02-10].http://www.linkedlifedata.com/.

3 Noor A,Assiri A,Ayvaz S,et al.Drug-drug interaction discovery and demystification using Semantic Web technologies[J].Journal of the American Medical Informatics Association Jamia,2017,24(3):556-564.

4 Huang J,Townsend C,Dou D,et al.OMIT:A Domain-Specific Knowledge Base for MicroRNA Tar-get Prediction[J]. Pharmaceutical Research,2011,28(12):3101-

3104.
5 Tenenbaum J D, Whetzel P L, Anderson K, et al. The Biomedical Resource Ontology (BRO) to enab-le resource discovery in clinical and translational research[J]. Journal of Bio-medical Informatics, 2011, 44(1): 137-145.
6 Kardan, A.A., Sani, M.F.& Modaberi, S. Implicit learner assessment based on semantic relevance of tags[J]. Computers in Human Behavior, 2016(55): 743-749.
7 张军亮.基于语义关联的多源医学信息资源发现服务系统研究[J].图书情报知识,2019(3):113-122.
8 张传文.基于大数据的区域医疗信息共享体系研究[D].广州:华南理工大学,2015.
9 杨帅旗.公众分类在医疗门户网站信息资源组织中的应用研究[D].北京:北京交通大学,2017.
10 ISO/IEC 13250:2003 Information technology-GML applications topic maps[EB/OL].[2020-03-17].Htt-ps://www.iso.org/obp/ui/#iso:std:iso-iec:13250:ed-2:v1:en.
11 Pepper S. The TAO of Topic Maps—Finding the way in the age of infoglut[EB/OL].[2020-03-17].http://www.ontopia.net/topicmaps/materials/tao.html.
12 施祷,熊回香,陆颖颖.基于主题图的非物质文化遗产数字资源整合实证分析[J].图书情报工作,2018,62(7):104-110.
13 Ontopia. Ontopia-the product[EB/OL].[2019-12-15].http://www.ontopia.net/section.jsp?id=Ontopia-the-p-roduct.
14 Rath H H. Topic Maps: Templates, Topology, and Type Hierarchies[J]. Acoustics Speech & Signal Processing Newsletter IEEE, 2000(2): 45-64.
15 Wang Y L, Wu B, Hu J Z. A Semantic Knowledge Fusion Method Based on Topic Maps[C]// Workshop on Intelligent Information Technology Application (IITA 2007). IEEE, 2008.
16 Robert C, Andres F, Veltman K. Advances in collaborative annotation in semantic management e-nvironment[C]// Digital Information Management, 2007. ICDIM ´07.2nd International Conference on. IEEE, 2007.
17 王平,吴玉萍.基于主题图的数字图书馆知识组织模型研究[J].情报理论与实践,2010(10):86-91.
18 李纲,王忠义.企业隐性知识管理方法研究[J].图书情报工作,2011,27(10):199-201.
19 夏立新,徐晨琛.基于主题图的电子政务门户知识导航系统构建研究[J].图书馆论坛,2010,30(6):184-187.

(责任编辑:徐　波)

Semantic Mining of Health Information Tag Based on Topic Map

ZHAN Ci, XIONG Hui-xiang, JIANG Wu-xuan, LI Yan

(School of Information Management, Central China Normal University, Wuhan 430079*, China)*

Abstract:【Purpose/significance】The effective organization of online health information is of great social value in improving the physical fitness of the whole people.【Method/process】Based on the analysis of health information topic, association relationship and resource indexing, an online health information tag semantic mining model based on the topic graph was constructed, and the health information tag topic graph was visualized.【Result/conclusion】Based on topic maps of online health information label semantic mining model can accurately find the deep relationship between online health information and label, can better reveal the semantic relationships of online health information tags to provide users with information visualization browsing and navigation to promote health information organization effect, help users to health information.【Innovation/limitation】This paper combines theme map with health information label to improve the efficiency of health information retrieval and utilization, but there are also shortcomings in this paper, such as small label sample size and sample range, lack of participation of professional medical researchers.

Keywords:health information; label; semantic mining; topic map; ontopia