

面向误诊提示的疾病-症状语义网构建

黄 岚^{1,2}, 纪林影³, 姚 刚⁴, 翟睿峰⁵, 白 天^{1,2}

(1. 吉林大学 计算机科学与技术学院,长春 130012; 2. 吉林大学 符号计算与知识工程教育部重点实验室,长春 130012; 3. 吉林大学 软件学院,长春 130012; 4. 吉林大学第二医院 神经内科,长春 130041; 5. 长春理工大学 电子信息工程学院,长春 130022)

摘要:首先,从医疗语料库中识别症状描述词语,基于症状间语义关系构建症状本体。然后,通过文本挖掘抽取疾病与症状间的关系、疾病与疾病的易误诊关系,建立疾病-症状语义网 DSSN。DSSN 中包含了疾病本体 DO、新构建的症状本体、疾病的易误诊关系及鉴别诊断知识。最后,通过一个临床诊断中的用例来说明 DSSN 在临床辅助诊断系统中对易误诊提示的帮助。

关键词:人工智能;语义网;文本挖掘;误诊;本体

中图分类号:TP182 **文献标志码:**A **文章编号:**1671-5497(2018)03-0859-07

DOI:10.13229/j.cnki.jdxbgxb20170406

Construction of disease-symptom semantic net for misdiagnosis prompt

HUANG Lan^{1,2}, JI Lin-ying³, YAO Gang⁴, ZHAI Rui-feng⁵, BAI Tian^{1,2}

(1. College of Computer Science and Technology, Jilin University, Changchun 130012, China; 2. Key Laboratory of Symbol Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun 130012, China; 3. College of Software, Jilin University, Changchun 130012, China; 4. Neurological Department, The Second Hospital of Jilin University, Changchun 130041, China; 5. College of Electronical and Information Engineering, Changchun University of Science and Technology, Changchun 130022, China)

Abstract: A Disease-Symptom Semantic Net (DSSN) for misdiagnosis prompt is constructed. First, symptom words are recognized from medical corpus, and a symptom ontology based on semantic relations between symptom words is established. Then, the relations between diseases and symptoms and the misdiagnosed relations between diseases were test mined and extracted to construct DSSN. DSSN contains Disease Ontology (DO), new established symptom ontology, misdiagnosed relations between diseases and differential diagnosis knowledge. Finally, a use case in clinical diagnosis is used to illustrate that DSSN is helpful to prompt misdiagnosis in clinical assistance diagnosis system.

Key words:artificial intelligence; semantic network; text mining; misdiagnosis; ontology

收稿日期:2017-04-21.

基金项目:国家自然科学基金项目(61472159, 61702214, 61572227);吉林省重点科技攻关项目(20160204022GX, 20170101006JC, 20170203002GX, 20150520064JH);吉林省产业创新专项项目(2017C030-1, 2017C033);中国博士后科学基金面上项目(2014M561293);珠海市优势学科项目;广东省优势重点学科建设项目。

作者简介:黄岚(1974-),女,教授,博士生导师。研究方向:社区发现,数据挖掘,商务智能。

E-mail:huanglan@jlu.edu.cn

通信作者:白天(1983-),男,副教授,博士。研究方向:生物信息学。E-mail:baitian@jlu.edu.cn

0 引言

误诊^[1]是临床诊疗中的一种常见现象。它造成的后果程度不一,轻者增加病人身心痛苦,延迟康复时间,重则危及生命,是医疗事故和医疗纠纷的主要原因之一。在临床诊断过程中,由于人们认识水平的局限性和疾病变化的复杂性,医生的初诊结果与疾病的实质不相符的现象时有发生,随着科学技术的进步和现代医学的发展,临床中不断引入各种现代化的检查仪器,使得诊断手段有了很大进步,然而临床误诊率并没有因此下降^[2]。根据粗略统计,疾病误诊率仍为 10%~15%^[3]。

误诊的最主要原因是相似症状的混淆。症状是临床诊断的主要依据,容易误诊的疾病通常是由其症状相似。易误诊疾病及症状的知识大量存储在各种书籍文献和开放的网络数据库中。因此,整合相关知识源,构建一个“疾病-症状”知识系统对疾病诊断过程中可能发生的误诊作出提示,对提升临床诊断效果有重要意义。

近年来,生物医学知识表示领域取得了一系列的进展:①结构化生物医学知识的表示及发现。本体是一种重要的结构化知识表示方法,是共享概念模型的明确的形式化规范说明,它的主要功能是实现知识的共享和复用^[4]。一些主要领域的本体已经建立,如基因本体^[5]、疾病本体^[6]、人类表型本体^[7]。②非结构化生物医学知识的表示及发现。近些年,大量生物医学信息和知识以学术论文、医学教科书、病例报告等半结构化和非结构化表示形式在互联网上发表。刘彦斌等^[8]针对语义生物信息库整合领域,解决了数据资源的链接问题。Mohammed 等^[9]通过构建疾病与症状之间的关系,将疾病本体与症状本体整合在一起。Cheng 等^[10]通过建立疾病相关数据库的语义关系来整合关于人类疾病的多种知识源。Huang 等^[11]设计了一种基于网络的算法,从多种生物医学语料库中抽取了疾病与基因的关系。Bai 等^[12]通过连接多种生物医学本体与知识源构建了一个混合的生物医学知识网络。然而,在对误诊提示的支持方面仍然存在一些尚未解决的问题。首先,现有的症状本体是基于解剖学的,其概念之间没有语义上的联系,使得症状间的相似关系在本体中没有得到体现。其次,症状与疾病间的关系存储在非结构化的文本中,未被抽取出来进行结

构化的表示。而且,症状与疾病间不是简单的一对一关系,还存在常见与罕见的区别。最重要的是,现有医学知识表示系统中都未包含疾病间的鉴别诊断(易误诊)知识。鉴别诊断知识通常存储在诊疗手册等文献中,尚未结构化地表达在计算机系统中,限制了疾病间易误诊知识的直接利用。

综上,本文构建了一个疾病-症状语义网(Disease-symptom semantic net, DSSN),其包含了疾病本体 DO、症状本体及疾病间的易误诊关系,并通过一个医学诊断中的例子来评估此语义网对于易误诊的提示作用。由于医学文献(如 PubMed 等)大多由英文表示,并且目前医学领域已建立的大量本体(如 DO)及各种术语标准(如 ICD-10 等)都是由英文表示的,因此本文构建的语义网也采用英文表示,并将在下一步工作中扩展中文版。

1 症状本体构建

构建症状本体的步骤为:①从 SYMP, Wikipedia, Cleveland Clinic 和 Mayo Clinic 等医疗领域知识库中获取描述症状文本的语料库;②在此语料库中进行症状描述词语的识别,得到尽量丰富的症状词汇候选集;③计算候选集中症状词汇间的语义相似度;④根据症状词语之间的语义相似度合并相近语义症状,建立一个新的症状本体。

1.1 语料库

为了获得全面、准确的症状词汇,本文选用 SYMP、Wikipedia、Cleveland Clinic 和 Mayo Clinic 作为语料库。SYMP(Symptom ontology)是一个以解剖学为基础的症状本体,包含 936 个症状;Wikipedia 中包含了以 ICD-10 为标准的全部疾病的百科知识页面,其中每个疾病都包含了其常见和罕见的症状描述;Cleveland Clinic 和 Mayo Clinic 是美国顶尖的综合医疗机构,集医疗服务、学术研究及教学于一体,其各自网站上都建立了疾病相关诊疗知识的知识库,包含各种疾病的症状描述。本文使用 SYMP 中的症状词汇,并分别提取 Wikipedia, Cleveland Clinic 和 Mayo Clinic 中关于疾病症状描述部分的文本作为症状识别的语料库。

1.2 症状词汇识别

生物医学注释工具,如 NCBO annotator^[13]和 MetaMap^[14],可以高精度地注释与疾病和症

状相关的术语。然而,在本文选用的语料库中,大量相关的词汇并不是出自现有疾病、症状等本体,甚至有些不是由单独的单词构成,所以不能被传统的注释工具完全识别。为了全面、准确地提取所有症状词汇,本文预先把文本语料库分为结构化语料(本体)和非结构化语料(文本)。对于这两种类型的语料,症状词汇识别流程为:①对于结构化语料,将之直接放入症状词汇候选集中;②对于非结构化语料,首先,使用 Porter Stemmer 算法^[15],在文本中提取与症状词汇候选集中词汇具有相同词根的词汇;然后,使用基于 WordNet^[16]的英语词汇相似度计算算法,进行特征提取,计算特征值,以此来提取文本中与候选集中词汇语义相似度高的词汇;最终,识别出文本中所有的症状词汇,图 1 为从 Wikipedia 中脑炎(encephalitis)的症状描述部分中识别出症状词汇。至此,本文就获得了扩充的症状词汇候选集,共为 2250 个。

Encephalitis
Signs and symptoms
Adult patients with encephalitis present with acute onset of fever, headache, confusion, and sometimes seizures. Younger children or Infants may present irritability, poor appetite and fever. Neurological examinations usually reveal a drowsy or confused patient. Stiff neck, due to the irritation of the meninges covering the brain, indicates that the patient has either meningitis or meningoencephalitis

图 1 症状词汇识别结果

Fig. 1 Result of symptom words recognition

1.3 同义症状合并

不同语料库对于同一种现象描述的词汇可能是不同的,例如对于“麻痹”这个症状,就有“paralysis”、“numbness”和“palsy”,为了统一表示相同意义的症状描述词汇,本文对症状描述词汇进行同义词合并处理。分为两步:①在症状语料中,对于同一症状有不同的描述,往往用括号注释,例如:difficulty swallowing (dysphagia),所以,首先识别文本中所有此类形式的同义词。②通过基于 WordNet 的英语词汇相似度计算算法^[16],计算词汇间相似度值,并以此识别词库中有着相同或相近语义的症状描述词汇,合并同义症状。

1.4 症状本体构建

已有的症状本体 SYMP 是基于解剖学构建的,例如把症状分为腹部症状、心血管系统症状、消化系统症状、神经系统症状和泌尿系统症状等,将症状与解剖学名词关联,如腹部绞痛、胸部充血。而在本文要建立的症状本体中,概念间的层

级构建是基于症状词汇间的语义关系。

前文已经进行了同义症状合并,得到了症状词汇的同义词(别称 Xref)。这些同义症状在所构建的症状本体中由一个概念(节点)表示。而表示相近症状描述的词语,根据其词根等词法特征以及语义范畴特征,建立这些相近症状词汇在语义上的“is-a”关系。例如:“痉挛”在症状本体中记为 spasm (cramp),其子类(近义词)为 muscle spasms, superimposed spasms 和 involuntary spasms。

2 疾病-症状语义网构建

2.1 疾病-症状关系抽取

症状的出现往往源自于某一疾病,疾病与症状之间的关联关系是临床诊断中的重要参考。疾病-症状关系抽取方法如下:首先,在 Wikipedia, Cleveland Clinic 和 Mayo Clinic 中获取描述某种疾病症状的文本;然后,基于扩充的症状词汇候选集,识别此段文本中的所有症状词语;最后,将提取的症状词汇关联到具体疾病,即建立症状-疾病间的关系“has symptom”。如图 1 所示,从 Wikipedia 中关于脑炎(Encephalitis)的症状描述文本中,提取了其中的症状词汇:acute onset of fever, headache, confusion, seizures, irritability, poor appetite, ever, drowsy, confused, stiff neck。从而,将脑炎与其症状连接起来。由此,建立所有疾病与症状的简单关系“has symptom”。

疾病与症状不完全是简单的一对一的关系,对于某种疾病,有些症状是常见的,有些症状是罕见的。本文除了建立“has symptom”这种疾病与症状间的简单关系外,将其常见或罕见程度的关系也提取并建立到疾病-症状语义网中。方法如下:首先在语料文本中获取描述频率的词语,本文确定了如表 1 所示的 10 余个描述症状频率的词汇。在语料文本中先定位到这些描述频率的词汇,再依据扩充的症状词汇候选集,提取同一句子中的症状词汇。依此可以确定某种疾病中该症状发生的频率,并将其建立到疾病与症状的关系中。

本文在 SYMP, Wikipedia, Mayo Clinic, Cleveland Clinic 中提取了 363 个频率词语,如表 1 所示。通过对临床医生的咨询,依据频率词语把症状分为 3 类:“most”,“most common”,“common”,“usually”,“often”,“≥10%”描述的症状是“常见症状”;“sometimes”,“less

“commonly”, “less often”, “3%~10%”描述的症状是“一般症状”; “occasionally”, “rare”, “≤3%”描述的症状是“罕见症状”。

表 1 频率提取结果

Table 1 Results of frequency extraction

症状	发生频率	总计
most	34	
most common	32	
common	56	
usually	54	328
often	115	
≥10%	37	
sometimes	17	
less commonly	2	23
less often	2	
3%~10%	2	
occasionally	6	
rare	4	12
≤3%	2	
总计	363	363

2.2 疾病间易误诊关系的获取及建立

疾病间的易误诊(鉴别诊断)知识,是构建疾病-症状语义网的核心。对疾病的误诊使很多本

不严重的病情,因为延误了治疗时机而带来了不良后果,易误诊知识对于疾病的诊断是必不可少的,所以本文在疾病-症状语义网中涵盖了这种知识。

对于疾病易误诊关系的获取,本文选用经典诊疗手册《Current essentials of medicine》^[17]作为知识源。书中描述了 561 种常见疾病的“诊断要点”和“鉴别诊断”信息,本文以此建立疾病间的易误诊关系。再将这种关系及鉴别诊断信息建立到疾病-症状语义网中,使得此语义网中的疾病根据易误诊关系相互关联,构成了在医疗诊断中进行误诊提示的知识库。如图 2 所示,以偏头痛(migraine)为例,首先,在书中获取关于偏头痛的页面,依据“鉴别诊断”文本,获取其易误诊疾病为:丛集性头痛(cluster headache)、脑膜炎(meningitis)、蛛网膜下腔出血(subarachnoid hemorrhage)等疾病;并依据“诊断要点”文本,获取到其鉴别诊断的描述性信息;最后,将此易误诊关系及鉴别诊断信息建立到疾病-症状语义网中。

Migraine headache

■ Essentials of diagnosis

- Onset usually in adolescence or early adulthood
- May be triggered by stress, foods(chocolate, red wine), smells (eg, perfume, car exhaust), dehydration, lack of sleep, menses
- Common migraine: Lasts 4~72 hours, unilateral, throbbing, moderate to severe intensity, aggravated by routine physical activity, associated with nausea, vomiting, photophobia, phonophobia
- Classic migraine (only approximately 20% of cases): Same symptoms as common migraine with a prodrome (aura) that includes a homonymous visual disturbance, unilateral numbness, paresthesias, or weakness
- Basilar variant: Brainstem and cerebellar findings followed by occipital headache
- Ophthalmic variant: Painless loss of vision, scotomas, usually unilateral

■ Differential diagnosis

- Cluster headache or other trigeminal autonomic cephalgia
- Giant cell arteritis
- Subarachnoid hemorrhage
- Mass lesion (eg, tumor or abscess)
- Meningitis
- Increased intracranial pressure of other cause

■ Treatment

- Avoidance of triggers
- Acute treatment: Triptans, ergotamine with caffeine, NSAIDs (preferably at onset of prodrome)
- Prophylaxis should be considered for more than three migraines per month and includes propranolol, amitriptyline, verapamil, valproic acid, and many others

Differential diagnosis: cluster headaches presents with one-sided nose stuffiness, tears and severe pain around the orbits, meningitis with fevers, and subarachnoid hemorrhage ...

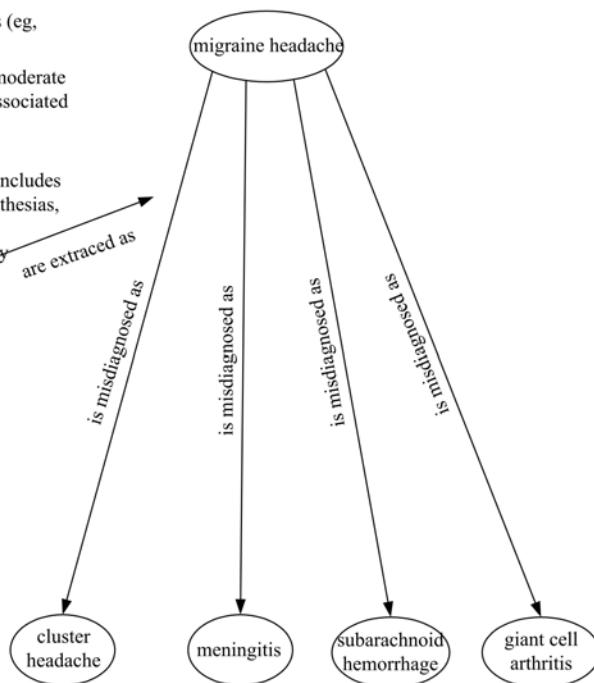


图 2 偏头痛易误诊疾病间关系,鉴别诊断知识的获取及建立

Fig. 2 Acquisition and construction of misdiagnosed diseases and differential diagnosis of migraine headache

2.3 疾病-症状语义网的建立

本文使用 Protégé^[18]构建了一个包含疾病本体、症状本体以及疾病之间易误诊关系的疾病-症状语义网。

OWL 类定义:类为疾病、症状词汇。

OWL 对象属性定义:对象属性代表类之间的关系,本文疾病与症状间对象属性分为“has symptom(A)”、“has symptom: common(B)”、“has symptom: general(C)”、“has symptom: rare(D)”以及疾病间的“is misdiagnosed as(E)”(易误诊)共5种,如表2所示。对于每一种对象属性,都有特定的范围和领域。

根据以上OWL定义,本文基于语义概念及其关系建立了疾病-症状语义网,其中有965个疾病词汇,2250个症状词汇,共3215个概念。此语

表2 对象属性描述

Table 2 Description of object properties

对象属性	范围和领域
A	Domain: “Disease” and Range: “Symptom”
B	Domain: “Disease” and Range: “Symptom”
C	Domain: “Disease” and Range: “Symptom”
D	Domain: “Disease” and Range: “Symptom”
E	Domain: “Disease” and Range: “Disease”

义网中,含有疾病-症状间的关系、疾病-疾病间的易误诊关系及鉴别诊断知识,所建立的DSSN的部分网络如图3所示。该语义网中表示了疾病之间的易误诊关系,以及易误诊疾病之间症状的异同。图4为易误诊疾病偏头痛(migraine headache)和丛集性头痛(cluster headache),以及两者间的各自症状区别。

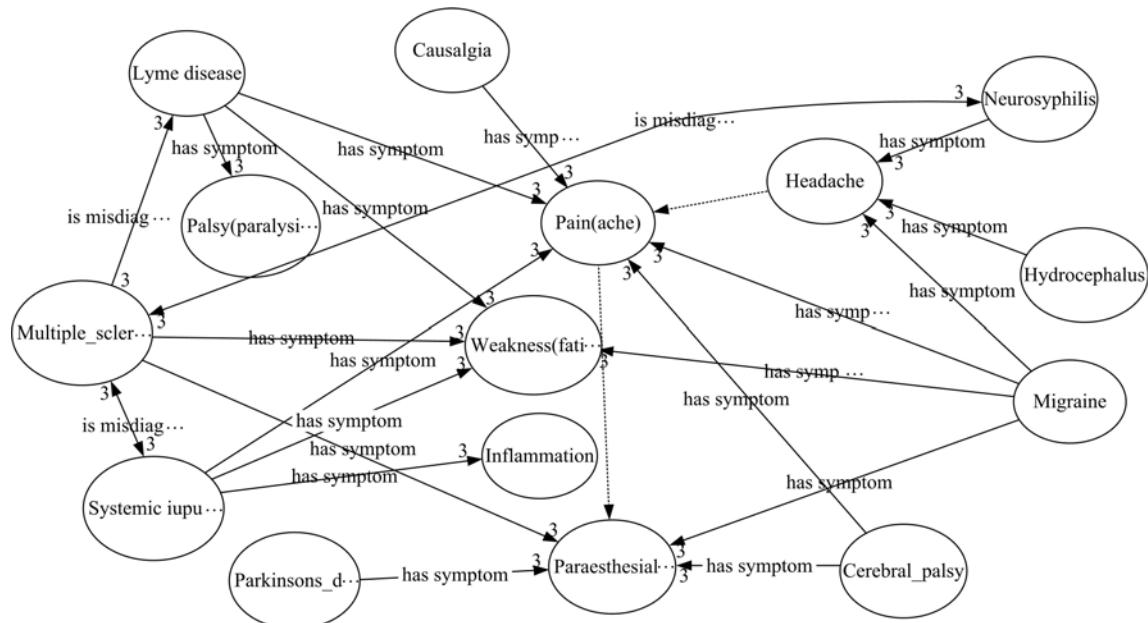


图3 疾病-症状语义网

Fig. 3 Disease-symptom semantic net

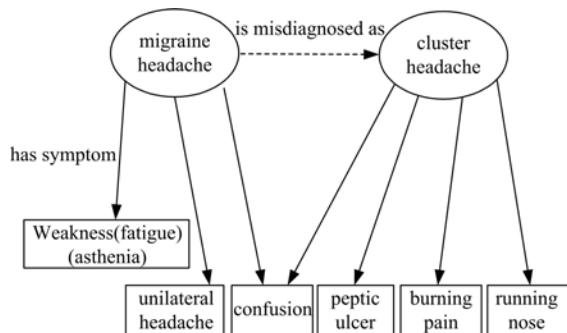


图4 DSSN中易误诊疾病例子

Fig. 4 Example of misdiagnosed disease in DSSN

3 DSSN 在误诊提示中的应用

相似症状的混淆是误诊发生的主要原因,鉴别诊断是防止和减少误诊发生的关键。为了快速、准确地对疾病进行区分和提示,必须要清晰地知道易误诊疾病间的不同症状和鉴别诊断知识。基于DSSN中疾病-症状间的关系、疾病-疾病间易误诊关系及鉴别诊断知识,可以对医疗诊断中的误诊进行提示。

阑尾炎(appendicitis)是最常见的外科急腹症之一,在临床实践中往往会出现误诊、误治现象,

其误诊率约为 9%~36%，大约 15% 的阑尾炎手术为误切正常阑尾^[19]。本文以阑尾炎为例，说明 DSSN 如何支持医疗诊断中的误诊提示。

基于 DSSN，可以在临床辅助诊疗系统中加入误诊提示模块。下面以一个用例来说明误诊提示模块的应用。假设医生对某位患者的初步诊断为阑尾炎，为防止误诊发生，他可以使用误诊提示模块在 DSSN 中通过检索阑尾炎相关易误诊知识来降低误诊发生的概率。如图 5 所示，在 DSSN 中获取阑尾炎的症状为右下腹痛、发烧、呕吐、便秘和脐周痛，其易误诊疾病为胃肠炎、胰腺炎、胆囊炎和异位妊娠。以上所列症状和易误诊

疾病对临床医生的诊断是一个提示，在所列出的易误诊疾病中，若医生认为胃肠炎也有潜在可能，他可以通过点击“胃肠炎”来返回当前初步诊断疾病“阑尾炎”与“胃肠炎”在症状上的异同，即阑尾炎与胃肠炎的相同症状及独有症状，如图 6 所示。它们相同症状为腹痛、呕吐、发烧，阑尾炎的独有症状为便秘、右下腹痛和脐周痛，胃肠炎的独有症状为腹泻和弥漫性疼痛。依此，医生就可以根据患者的具体情况作出鉴别诊断。综上，基于 DSSN 的易误诊模块，能够可视化地展现疾病及其易误诊疾病的症状，对医生在医疗诊断中进行误诊提示，从而降低误诊发生的概率。

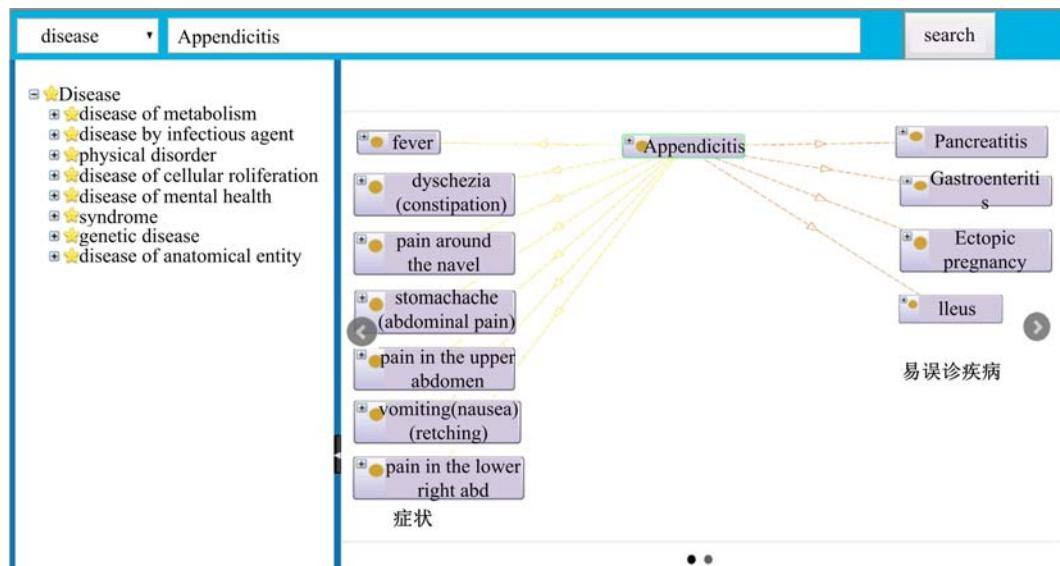


图 5 阑尾炎的症状及其易误诊疾病

Fig. 5 Symptoms of appendicitis and its misdiagnosed diseases

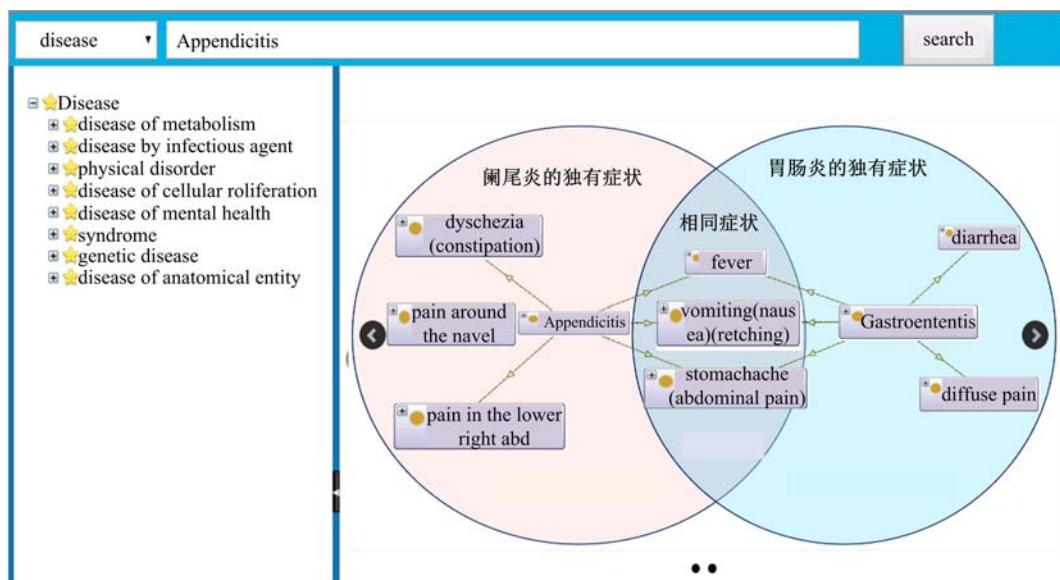


图 6 阑尾炎和胃肠炎的相同及独有症状

Fig. 6 Same symptoms and unique symptoms between appendicitis and gastroenteritis

4 结束语

临床诊断中误诊现象十分普遍,其主要原因是易误诊疾病之间有着相似的症状。而疾病之间的易误诊关系,以及不同疾病之间症状的区别等知识在各种医疗文献中已经大量存储。本文通过构建一种疾病-症状语义网(DSSN),将这些知识结构化地表达在语义网中,即通过对多个医学领域知识库进行自然语言处理和文本挖掘,获得扩充的症状词汇候选集、疾病-症状间的关系以及疾病间的易误诊关系和知识,并将这些关系和知识建立成语义网的表达形式。此外,本文还通过一个用例说明了构建的DSSN在临床辅助诊断系统中对易误诊疾病提示的帮助。

参考文献:

- [1] Balogh E P, Miller B T, Ball J R, et al. Improving diagnosis in healthcare [R]. Washington, DC: The National Academies Press, 2015.
- [2] 赵会懂. 通过误诊文献的关键词词频分析看临床误诊发生的规律[J]. 中国医学图书情报杂志, 2010, 19(11): 73-77.
Zhao Hui-dong. Clinical misdiagnosis rules found in analysis of the frequencies of its key words appeared in misdiagnosis related papers [J]. Chinese Journal of Medical Library and Information Science, 2010, 19(11): 73-77.
- [3] Ely J W, Graber M L, Croskerry P. Checklists to reduce diagnostic errors[J]. Academic Medicine, 2011, 86(3): 307-313.
- [4] 王俊华,左万利,彭涛. 面向文本的本体学习方法[J]. 吉林大学学报:工学版, 2015, 45(1): 237-244.
Wang Jun-hua, Zuo Wan-li, Peng Tao. Test-oriented ontology learning methods[J]. Journal of Jilin University (Engineering and Technology Edition), 2015, 45(1): 237-244.
- [5] Gene O C. The gene ontology (GO) database and informatics resource[J]. Nucleic Acids Research, 2004, 32(Database): D258-D261.
- [6] Schriml L M, Arze C, Nadendla S, et al. Disease ontology: a backbone for disease semantic integration[J]. Nucleic Acids Research, 2012, 40 (Database): D940-D946.
- [7] Kohler S, Doelken S C, Mungall C J, et al. The human phenotype ontology project: linking molecular biology and disease through phenotype data[J]. Nucleic Acids Research, 2014, 42(Database): D966-D974.
- [8] 刘彦斌,周春光,张重毅,等. 基于本体的语义生物信息数据库资源链接[J]. 吉林大学学报:工学版, 2010, 40(6): 1651-1654.
Liu Yan-bin, Zhou Chun-guang, Zhang Chong-yi, et al. Resource interlinking of semantic bioinformatics database based on ontology[J]. Journal of Jilin University (Engineering and Technology Edition), 2010, 40 (6): 1651-1654.
- [9] Mohammed O, Benlamri R, Fong S. Building a diseases symptoms ontology for medical diagnosis: an integrative approach[C]// Proc of the IEEE International Conference on Future Generation Communication Technology, London, 2012: 104-108.
- [10] Cheng L, Wang G H, Li J, et al. SIDD: a semantically integrated database towards a global view of human disease[J]. Plos One, 2013, 8(10): e75504.
- [11] Huang Lan, Wang Ye, Wang Yan, et al. Gene-disease interaction retrieval from multiple sources: a network based method[J]. Biomed Research International, 2016, 2016(3): 3594517.
- [12] Bai Tian, Gong Lei-guang, Wang Ye, et al. A method for exploring implicit concept relatedness in biomedical knowledge network[J]. BMC Bioinformatics, 2016, 17(9): 53-66.
- [13] Jonquet C, Shah N H, Youn C H, et al. NCBO annotator: semantic annotation of biomedical data[C]// Proc of the 8th Int Semantic Web Conf, Poster and Demonstration Session, Berlin, 2009: 2-3.
- [14] Demner-Fushman D, Rogers W J, Aronson A R. MetaMap lite: an evaluation of a new Java implementation of MetaMap[J]. Journal of the American Medical Informatics Association, 2017, 24(4): 841-844.
- [15] Porter M F. An algorithm for suffix stripping[J]. Program Electronic Library & Information Systems, 2006, 14(3): 130-137.
- [16] Miller G A, Fellbaum C. WordNet then and now[J]. Language Resources and Evaluation, 2007, 41(2): 209-214.
- [17] Lawrence Tierney, Saint Sanjay, Whooley Mary. Current Essentials of Medicine[M]. 4th ed. New York: McGraw-Hill Companies, 2011.
- [18] Gennari J H, Musen M A, Fergerson R W, et al. The evolution of Protégé: an environment for knowledge based systems development[J]. International Journal of Human-Computer Studies, 2009, 58(1): 89-123.
- [19] 孙宝志. 临床医学导论[M]. 北京:高等教育出版社, 2013.