基于语义谓词优化的疾病因果关系发现*

董鹏 李晓瑛 李军莲 唐小利
（北京协和医学院/中国医学科学院/医学信息研究所/图书馆，北京 100005）

摘要：挖掘PubMed生物医学文献中蕴含的疾病因果关系，能够为疾病预防、诊疗提供参考，促使建立更好的疾病预防和治疗措施。本研究提出基于SemRep语义谓词优化的疾病因果关系抽取方法，构建了包含50个语义谓词的疾病因果关系语义谓词表，自动抽取259 434条疾病因果关系对，揭示疾病之间的单向因果关系和双向因果关系，并结合可视化图形直观呈现。最终验证了优化语义谓词的方法提升SemRep对疾病因果关系抽取效果的可行性，能够有效地从大规模生物医学文献中抽取疾病因果关系，也为其他领域的语义关系精准发现提供借鉴。

关键词：语义谓词优化；疾病因果关系发现；SemRep
中图分类号：G203　DOI：10.3772/j.issn.1673-2286.2022.11.007
引文格式：董鹏，李晓瑛，李军莲，等. 基于语义谓词优化的疾病因果关系发现[J]. 数字图书馆论坛，2022（11）：19-25.

疾病与疾病之间存在因果关系，即当一种疾病发生时，可能会后继引发一种甚至一系列的疾病。揭示疾病因果关系，可以阐释疾病发生、发展的机理，并促使建立更好的疾病预防和治疗措施[1-2]。当前，研究人员已经使用基因组学、表型组学等科学数据，从分子层面开展疾病的归因研究，探索疾病之间的因果关系或其他关联关系。随着研究不断深入，生物医学文献数据受到研究人员的重视和应用。生物医学文献不但记载了科学实验的过程与结果，还记载了人类治疗疾病的临床诊疗经验，蕴含大量的疾病因果关系。准确、细粒度的揭示文献中蕴含的疾病因果关系，可以提升文献的利用价值，促进生物医学文献数据与科学数据的有效融合。

1 研究背景

从PubMed文献数据中抽取生物医学语义关系，可以用于揭示疾病、药物、蛋白质、基因等医学实体之间的关联关系，支撑临床和科研任务。在具有大量医学知识资源积累（叙词表、本体等）、规则构建精准的生物医学领域，基于规则的语义关系抽取方法具有良好的效果[3]。基于规则的方法主要借助已有知识积累，与共现分析、人工制定语义关系模板相结合抽取语义关系，原理简单、过程清晰、结果易懂。

SemRep[4-5]由美国国立医学图书馆（NLM）基于一体化医学语言系统（Unified Medical Language System，UMLS）[6]开发，结合了自然语言处理技术与UMLS中包含的结构化生物医学领域知识，使用基于规则的方法从PubMed文献数据中抽取种类广泛的生物医学语义关系，并以S-P-O三元组模式进行格式化表示与存储。其中，主语和宾语来自UMLS专家词典中的归一化名词概念；语义谓词来自UMLS语义网络中的58种规范语义关系类型[6]。SemRep以简单易用和高效的特点被广泛用于生物医学实体间的语义关系抽取和发现，如治疗/因果关系[7]、临床决策[8]、矛盾知识识别[9-10]等。

在已有研究中，SemRep表现为53%~83%的准确率和42%~53%的召回率，错误分析显示，语义谓词识

* 本研究得到中国医学科学院医学与健康科技创新工程重大协同创新项目“生物医学文献信息保障与集成服务平台”（编号：2021-I2M-1-033）资助。

别不准是导致SemRep抽取结果错误的重要原因[11-13]。SemRep通过自动筛选主语、宾语和规范化谓词实现语义关系自动抽取，然而具体有哪些文本语义谓词被SemRep识别、归并到“CAUSES”中，并且这些文本语义谓词表示疾病间因果关系的准确率如何，因其高度封装，使用者不得而知。基于此，本研究计划对SemRep识别、抽取出的文本语义谓词进行评估、筛选，通过优化语义谓词的方法来解决SemRep在抽取特定语义关系时语义谓词识别不准的问题，以期提升SemRep自动抽取疾病因果关系的准确率和文献中蕴含疾病因果关系的发现效果。

2 研究方法和总体思路

2.1 研究方法与数据来源

本研究通过语义分析和实验评估方法，在SemRep解析、识别出的语义谓词中筛选表达疾病间因果关系较为准确的语义谓词，实现语义谓词优化，提升SemRep自动抽取疾病因果关系的效果，进而在生物医学文献中自动抽取和发现疾病因果关系。本研究的基础数据来自SemMedDB（Semantic MEDLINE Database）[14]，这是一个大型语义关系数据库，以三元组结构化形式保存了SemRep工具对PubMed全部文献数据的语义关系解析结果，仅可用于非商业用途。

2.2 基于语义谓词优化的疾病因果关系发现总体思路

生物医学文献中的实体关系表达主要依赖自然语言中能够表示语义关系的语义谓词，通过这些语义谓词，不但可以确定实体间存在关联关系，还可以确定其关系类型，具有较好的关系揭示效果[15-16]。基于此，本研究通过对SemRep识别出的文本谓词进行评估、筛选，提升SemRep自动抽取疾病因果关系的效果，并发现生物医学文献中蕴含的疾病因果关系。

为实现研究目标，本研究设计了基于语义谓词优化的疾病因果关系发现总体思路（见图1），主要包括如下3个步骤。①对来源数据进行数据预处理，构建研究所需的数据集。②在参考谓词中提取语义特征词，并在基础数据集中获取更多的谓词模式。通过实验评估，在获取的全部谓词模式中筛选出表达疾病间因果关系较

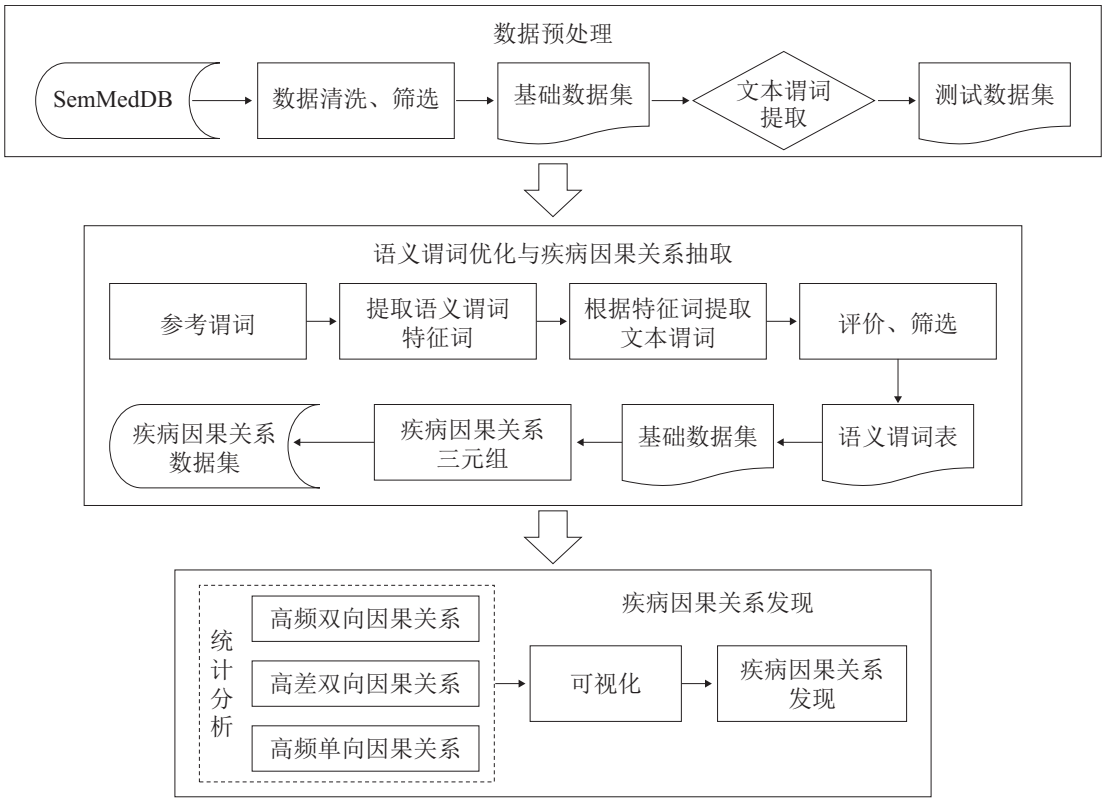

图1 基于语义谓词优化的疾病因果关系发现总体思路

为准确的语义谓词，完成语义谓词优化与疾病因果关系抽取。③分析、解读抽取的疾病因果关系，实现基于语义谓词优化的疾病因果关系发现。

3 研究过程及结果

3.1 数据预处理与SemRep评估

3.1.1 数据预处理

SemMedDB中的每条关系数据都含有丰富的字段内容[17]，除三元组外，还包括三元组相关句子、谓词坐标等信息。疾病关系数据筛选和整理步骤如下。

首先，通过对主语和宾语的语义类型筛选，只保留“SUBJECT_SEMTYPE”和“OBJECT_SEMTYPE”语义类型为“dsyn”（疾病）、“mobd”（精神疾病）、“neop”（肿瘤）的关系对；筛选规范化谓词“PREDICATE”，通过语义谓词“PREDICATE = ‘ISA’ ”，排除“属种关系”的疾病对。

其次，通过主语和宾语的NOVELTY属性值为1，排除泛指疾病名称，如“Disease”；同时发现NOVELTY属性值为0的情况下，仍有泛指疾病名称，如“Malignant Neoplasms”“Neoplasm”“Infection”等，进行剔除。另外，在数据检查中发现句子中的形容词“little”被识别为“Little’s Disease”（李特尔氏病，痉挛性大脑性两侧瘫），一并剔除。

最后，获得1 268 284条疾病关系对，将其保存在关系型数据库中，作为本研究的基础数据集。SemRep处理后的关系数据对语义谓词进行了自动归并，输出结果不包括文本谓词，本研究根据输出结果中的文本句和文本谓词起止坐标，使用SQL语言编程提取出了文本谓词。

3.1.2 SemRep疾病因果关系抽取评估

为评估SemRep自动抽取疾病因果关系的效果，本研究创建了小型测试数据集。在基础数据集中随机抽取500篇文献，获取其中的全部疾病关系对，数据的判别标注工作请2名专家背对背完成，根据文本句审核、判别关系对是否属于疾病因果关系，有疑异的标注结果由2名专家当面讨论后达成一致。最终测试数据集中有疾病关系对741个，其中疾病因果关系对304个。另外，对304个疾病因果关系对中的语义谓词进行分析和准确率评估，获得表达疾病间因果关系的语义谓词28个。

本研究使用了准确率（Precision）、召回率（Recall）以及综合评价指标F值（F-Measure）对实验结果进行评估。在测试数据集中，SemRep自动抽取疾病因果关系的准确率、召回率和F值分别为85.34%、53.62%、65.86%。人工审核中注意到，在某些情况下规范语义谓词“CAUSES”影响了SemRep自动抽取疾病因果关系的结果，即某些语义谓词未被归并到“CAUSES”中或被归并到“CAUSES”中的一些谓词不能准确表示疾病因果关系。因此，在基于SemRep解析结果的人工审核抽取中补充了部分语义谓词，并得到95.49%的准确率、83.55%的召回率和89.12%的F值。实验表明，SemRep工具可以支持疾病因果关系抽取工作，并可以通过筛选、优化语义谓词提升SemRep工具自动抽取疾病因果关系的性能。

3.2 语义谓词优化及疾病因果关系抽取

本研究中语义谓词优化的流程如下。

首先，在表达疾病因果关系的参考谓词中提取语义特征词。一部分参考谓词来自Xu等[18]揭示的26个高准确率的语义谓词；另一部分参考谓词是自本研究测试数据集中获取的28个语义谓词。汇总、去重后共获得参考谓词49个，这些谓词大多以词组形式存在，从中提取了49个参考谓词中的语义特征词，如参考谓词“due to”中的语义特征词“due”。

其次，根据语义特征词在基础数据集中筛选更多的谓词形式。在基础数据集中，包含语义特征词的文本谓词都会被提取。请2名专家对提取的谓词在“是否为合理的谓词形式”和“是否可以表达疾病因果关系”两个方面进行背对背的审核，有疑异的结果由专家当面讨论后达成一致。根据审核结果清洗文本谓词：剔除字节数超长、明显不合理谓词形式，如“heart disease and dementia were the risk factors of this disease”；剔除不能表达疾病因果关系的谓词形式，如“cause decrease”“cause a change”；剔除错误抽取的谓词形式，如根据语义特征词“owing”提取出的“following”。最终共获得可以表示疾病间因果关系的语义谓词56个。

再次，为定量揭示每个语义谓词表达疾病因果关系的准确率，在基础数据集中分别为每个语义谓词随

机抽取50个疾病因果关系对（部分谓词的关系对总数不足50条），按上一步骤中人工审核的方式评估每个语义谓词的准确率，其中准确率不低于80%的谓词有36个，准确率不低于60%的谓词有42个，准确率不低于40%的语义谓词50个，以这些语义谓词构建疾病因果关系语义谓词表，实现语义谓词优化。

最后，在测试数据集上检验基于语义谓词表自动抽取疾病因果关系的效果。将谓词表中的谓词按最低准确率分别为80%、60%、40%进行评估实验，分别使用了36个、42个和50个语义谓词。实验结果如表1所示，随着使用的语义谓词准确率下降，疾病因果关系抽取的准确率呈下降趋势，召回率和F值呈上升趋势。三次测试的准确率分别比未优化语义谓词的自动抽取结果（85.34%）提高了13.63%、12.74%和8.31%。通过实验，确定通过优化语义谓词提高SemRep自动抽取疾病因果关系准确率的方法可行。

表1 语义谓词表用于疾病因果关系抽取评估结果

谓词数量/个	准确率/%	召回率/%	F值/%
36	96.97	63.16	76.50
42	96.21	66.78	78.84
50	92.43	76.32	83.60

与Xu等[18]揭示的26个语义谓词相比，语义谓词表的谓词更多，覆盖更多疾病因果关系表示形式，可以从文本中抽取更多的疾病因果关系对。与SemRep的规范谓词相比，语义谓词表包括被归并到非“CAUSES”且可以表示疾病因果关系的语义谓词，如分别具有90%和80%准确率的“risk factor”和“led”被分别归一化为“PREDISPOSES”和“AFFECTS”；同时，发现SemRep归一化谓词“CAUSES”中个别语义谓词表示疾病间因果关系的准确率较低，如“because of”在测试中的准确率仅有36%，排除这些谓词有助于提高SemRep自动抽取疾病因果关系的效率。

3.3 疾病因果关系发现

疾病因果关系抽取中，使用语义谓词表中准确率不低于80%的36个语义谓词与基础数据集中的文本谓词自动匹配，只有两者完全匹配的疾病对才会被选中。最终共自动抽取259 434条疾病因果关系三元组，保存于关系型数据库中。

通过对抽取结果的统计分析，发现两种疾病间除存在一种疾病导致另一种疾病的单向因果关系外，还存在两种疾病互为病因的双向因果关系，例如文献（PMID：17438881）中表述“Endothelial dysfunction is an important factor leading to atherosclerosis，Hypertension and heart failure”，认为内皮功能障碍（Endothelial dysfunction）可以引发高血压（Hypertensive disease）；亦有文献（PMID：16715652）报道高血压可以引发内皮功能障碍，如“Hypertension causes Endothelial dysfunction”。这种情况在本研究中被认为是两种疾病间的双向因果关系，并将频次较高的疾病关系对作为正向因果关系、频次较低的关系对作为反向因果关系。

3.3.1 单向高频疾病因果关系

未经过人工审核的疾病因果关系自动抽取结果存在错误数据，为减少错误数据干扰，本研究利用关系对在文献中的共现频次进行筛选，认为频次不低于10次的单向疾病因果关系对可信度较高，最终筛选得到疾病因果关系对41 724对，涉及1 796种疾病。为便于展示，研究中仅对频次不低于100次的33种疾病因果关系进行可视化（见图2）。图中疾病左侧为“因”、右侧为“果”，连线粗细表示相应疾病因果关系的频次高低。

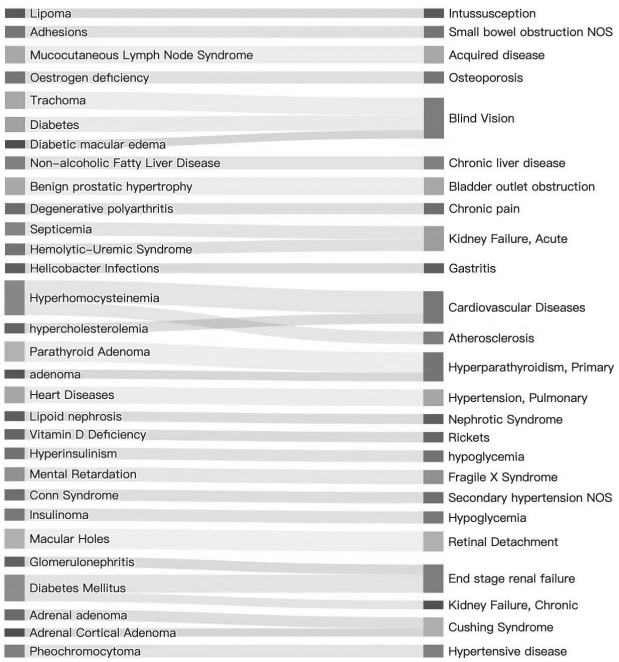

图2 单向疾病因果关系

图2中，与失明（Blind Vision）相关的疾病因果关系频次最高，揭示最大致盲因素来自沙眼（Trachoma）和糖尿病性黄斑水肿（Diabetic macular edema）等疾病，失明往往作为这些疾病不断进展的严重后果；其次揭示高同型半胱氨酸血症（Hyperhomocysteinemia）同时是心血管疾病（Cardiovascular Diseases）和动脉粥样硬化（Atherosclerosis）的致病因素。图2同样直观揭示引发急性肾衰竭（Kidney Failure，Acute）、慢性肾衰竭（Kidney Failure，Chronic）和终末期肾衰竭（End stage renal failure）的疾病。

3.3.2 双向疾病因果关系

本研究筛选了共现频次不低于10次的双向疾病因果关系对，认为这些关系对是双向高频疾病因果关系，共涉及56种疾病。作为双向高频疾病因果关系的示例，研究中绘制了肥胖症（Obesity）和高血压（Hypertensive disease）的加权双向疾病因果关系图（见图3）。其中权重由关系对在文献中的出现频次确定，连线方向表示疾病因果关系的方向，连线的粗细由正方频次与反向频次之和决定。

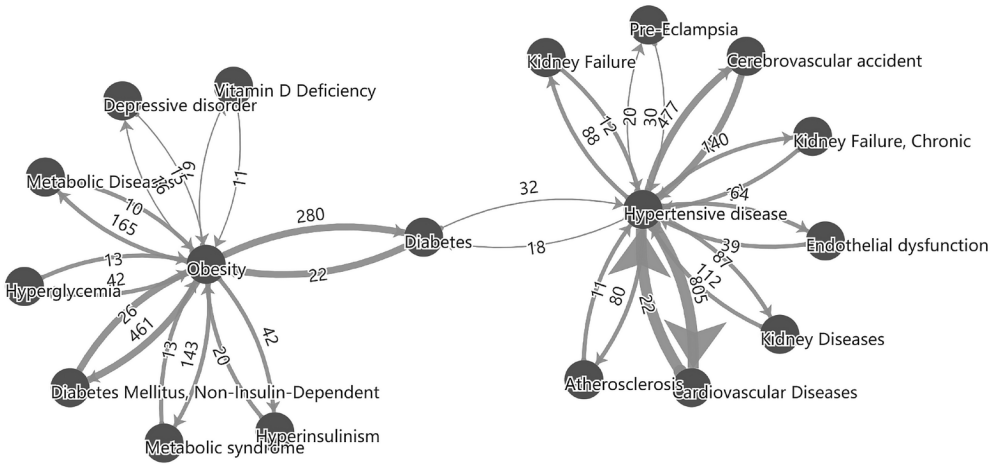

图3 肥胖症与高血压的加权双向疾病因果关系图

肥胖症作为一种常见的代谢病，已发展成为全球流行病。与肥胖症有双向高频因果关系的疾病有8种，肥胖症与非胰岛素依赖型糖尿病（Diabetes Mellitus，Non-Insulin-Dependent_Obesity）的因果关系最为密切，肥胖症可通过引发脂肪组织炎症导致胰岛素抵抗和长期的非胰岛素依赖型糖尿病，而反向关系的非胰岛素依赖型糖尿病导致肥胖症观点虽然被确切记载于文献中，但缺少可以支持这一观点的机理研究，这个发现可能为相关研究提供参考。

众所周知，高血压（Hypertensive disease）是心血管疾病（Cardiovascular Diseases）和脑血管意外（Cerebrovascular accident）的重要病因。在反向因果关系中，心血管疾病作为高血压的罕见病因，通过影响人体免疫系统或心血管异常而导致高血压；约75%的患者会因为脑血管意外（中风或卒中）导致中风性高血压。这些因果关系可以为疾病的临床诊疗提供参考。

在进一步的分析中发现，一些疾病间的正向与反向因果关系的共现频次存在较大差异，本研究将这种情况称为双向高频差疾病因果关系，并通过计算方法量化$D=(F-N)/N$，D为正反向频次的频差值，F表示正向频次，N表示反向频次。研究中分析了反向频次低于10且频差大于10的双向高频差疾病因果关系，以期发现两种疾病间因果关系的不确定性，为临床、科研提供研究方向。排除错误数据后，表2展示了前10组双向高频差的疾病因果关系。

结合关系对出处文献，对表2中的双向疾病因果关系进行分析，可以发现疾病因果关系的不确定性。如阻塞性睡眠呼吸暂停（Sleep Apnea，Obstructive）是一种常见的睡眠障碍，肥胖症（Obesity）已被明确是引发阻塞性睡眠呼吸暂停的主要危险因素，但对于阻塞性睡眠呼吸暂停导致肥胖症的研究成果较少，有待研究人员进行更多研究。另外，大量流行病学文献已经证实，哮喘（Asthma）与肥胖症（Obesity）互为因果关系，但两者之间相互引发的作用机制被认为尚不够深入和明确。

表2 双向高频差疾病因果关系对

疾病因果关系对			F/次	N/次	D
序 号	因	果			
1	Coronary Artery Vasospasm	Myocardial Infarction	97	3	31.33
2	Hypertension, Pulmonary	Right ventricular failure	127	4	30.75
3	Obesity	Fatty Liver	105	4	25.25
4	Ischemia	Ventricular Fibrillation	158	7	21.57
5	Hypertensive disease	End stage renal failure	180	8	21.5
6	Subarachnoid Hemorrhage	Cerebral Vasospasm	152	7	20.71
7	Obesity	Asthma	151	7	20.57
8	Obesity	Sleep Apnea, Obstructive	95	5	18
9	Bacterial Infections	Septicemia	69	5	12.8
10	Strabismus	Amblyopia	63	5	11.6

4 总结与展望

本研究主要完成两部分工作：①评估、筛选SemRep识别出的文本语义谓词，获取表达疾病因果关系准确率较高的谓词模式，构建疾病因果关系语义谓词表，实现语义谓词优化，提升SemRep自动抽取疾病因果关系的效果。②基于语义谓词表自动抽取疾病因果关系，发现生物医学文献中的疾病因果关系。研究意义在于从语义层面细粒度地揭示生物医学文献中的特定语义关系，可以提升生物医学研究人员对大规模生物医学文献的利用效率，有助于探索更佳的临床治疗方案和疾病防控机制。此外，以S-P-O三元组形式提供机器可理解、可计算、可推理的结构化疾病因果关系数据，有助于促进生物医学文献数据与科学数据在语义层面的有效融合，为进一步探索疾病间的潜在因果关系、提出疾病因果关系假设，提供良好的数据基础，助力疾病“归因研究”。对应用SemRep自动发现医学实体间特定语义关系等相关研究，本文所使用的方法具有可移植性和适用性。在后续研究中，将探索、改进语义谓词的优化方法和流程，在充分利用现有丰富医学知识和语义规则的基础上，实现大规模生物医学文献中的疾病因果关系精准发现。

参考文献

[1] BACH J F. Causality in medicine [J]. Comptes Rendus Biologies, 2019, 342 (3/4): 55-57.

[2] 徐静汶，李晓彬，王学习，等. 疾病因果逻辑关系的辩证思维 [J]. 医学与哲学，2013，34 (6)：1-3，18.

[3] 李芳，刘胜宇，刘峥. 生物医学语义关系抽取方法综述 [J]. 图书馆论坛，2017，37 (6)：61-69.

[4] RINDFLESCH T C, FISZMAN M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text [J]. Journal of Biomedical Informatics, 2003, 36 (6): 462-477.

[5] 丁云轩，闫雷. 数据挖掘软件SemRepr的评价 [J]. 中华医学图书情报杂志，2008，17 (6)：71-75.

[6] 李晓瑛，李军莲，李丹亚. 一体化医学语言系统及其在知识发现中的应用研究 [J]. 数字图书馆论坛，2019，4 (9)：24-29.

[7] BAKAL G, TALARI P, KAKANI E V, et al. Exploiting semantic patterns over biomedical knowledge graphs for predicting treatment and causative relations [J]. Journal of Biomedical Informatics, 2018, 82: 189-199.

[8] MORID M A, FISZMAN M, RAJA K, et al. Classification of clinically useful sentences in clinical evidence resources [J]. Journal of Biomedical Informatics, 2016, 60: 14-22.

[9] ROSEMBLAT G, FISZMAN M, SHIN D, et al. Towards a characterization of apparent contradictions in the biomedical literature using context analysis [J]. Journal of Biomedical Informatics, 2019, 98: 103275.

[10] 王雪，杨雪梅，李沛鑫，等. 基于语义模型的药物矛盾知识发现 [J]. 情报杂志，2020，39 (7)：159-165.

[11] AHLERS CB, FISZMAN M, DEMNER-FUSSHMAN D, et al. Extracting semantic predications from Medline citations for pharmacogenomics [C] //Pacific Symposium on

Biocomputing2007. Hackensack：World Scientific，2007：209-220.

[12] HRISTOVSKI D，DINEVSKI D，KASTRIN A，et al. Biomedical question answering using semantic relations [J]. BMC Bioinformatics，2015，16：6.

[13] KILICOGLU H，ROSEMBLAT G，FISZMAN M，et al. Broad-coverage biomedical relation extraction with SemRep [J]. BMC Bioinformatics，2020，21：188.

[14] KILICOGLU H，SHIN D，FISZMAN M，et al. SemMedDB：a PubMed-scale repository of biomedical semantic predications [J]. Bioinformatics，2012，28（23）：3158-3160.

[15] 王秀艳，崔雷. 应用关键动词抽取生物医学实体间语义关系研究综述 [J]. 现代图书情报技术，2011（9）：21-27.

[16] KILICOGLU H，ROSEMBLAT G，FISZMAN M，et al. Constructing a semantic predication gold standard from the biomedical literature [J]. BMC Bioinformatics，2011，12：486.

[17] NLM. SemMedDB Database Details [EB/OL]. [2022-09-19]. https://lhncbc.nlm.nih.gov/ii/tools/SemRep_SemMedDB_SKR/dbinfo.html.

[18] XU R，LI L，WANG Q. dRiskKB：a large-scale disease-disease risk relationship knowledge base constructed from biomedical text [J]. BMC Bioinformatics，2014，15：105.

作者简介

董鹏，男，1986年生，硕士研究生，馆员，研究方向：医学知识组织与知识发现。
李晓瑛，女，1982年生，博士，副研究员，研究方向：医学知识组织与知识发现。
李军莲，女，1972年生，博士，研究馆员，研究方向：医学知识组织与信息处理。
唐小利，女，1966年生，硕士，研究馆员，通信作者，研究方向：医学信息服务与情报分析，E-mail：tang.xiaoli@imicams.ac.cn。

Disease Causality Discovery Based on Semantic Predicates Optimization

DONG Peng　LI XiaoYing　LI JunLian　TANG XiaoLi
(Institute of Medical Information/Medical Library, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100005, P. R. China)

Abstract: Exploring the causality of disease in Pubmed biomedical literature can provide reference for prevention, diagnosis and treatment of disease, further improve relative measure for disease prevention and treatment. This study proposes a disease causality extraction method based on SemRep semantic predicate optimization, constructs a disease causal relationship semantic predicate table containing 50 semantic predicates, automatically extracts 259 434 disease causal relationship pairs, reveals the one-way causal relationship and two-way causal relationship between diseases, and visually presents them with visual graphics. Finally, the feasibility of optimizing semantic predicates to improve the effect of SemRep on disease causal relationship extraction is verified, which can effectively extract disease causal relationship from large-scale biomedical literature, and also provide reference for accurate discovery of semantic relationship in other fields.

keywords: Semantic Predicates Optimization; Disease Causality Discovery; SemRep

（收稿日期：2022-10-20）