

基于文献的生物医学知识发现研究综述

郎宇翔 杨艳萍

(1 中国科学院文献情报中心 北京 100190 2 中国科学院大学图书情报与档案管理系 北京 100049)

〔摘要〕 介绍基于文献的知识发现 (Literature Based Discovery, LBD) 概念及实现方法, 从数据获取、概念抽取、关系挖掘、结果评估 4 个方面阐述 LBD 相关研究情况及关键问题, 分析 LBD 在生物医学领域的应用情况及 LBD 研究存在的问题, 提出建议和展望。

〔关键词〕 文献知识发现; 潜在关联发现; 文本挖掘

〔中图分类号〕 R-058 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2021.10.007

Review on Literature Based Discovery Researches in Biomedical Field LANG Yuxiang, YANG Yanping, 1National Science Library, Chinese Academy of Sciences, Beijing 100190, 2Department of Library & Information Science and Archives Management, College of Economic and Management, University of Chinese Academy of Sciences, Beijing 100049, China

〔Abstract〕 The paper introduces the concept and implementation method of Literature Based Discovery (LBD), elaborates the research situation and key issues related to LBD from four aspects of data acquisition, concept extraction, relationship mining and result evaluation, analyzes the application situation of LBD in biomedical field and the existing problems of LBD researches, and puts forward some suggestions and prospects.

〔Keywords〕 Literature Based Discovery (LBD); latent relationship discovery; text mining

1 引言

随着生物医学领域研究发展, 相关文献数量呈指数级增长, 仅 PubMed 数据库中的文献数量就已超过 3 000 万^[1], 科研人员利用文献进行科学研究的难度提升。数据密集型研究范式强调用数据驱动科学发现, 生物医学领域研究文献包含大量隐含关联, 对其进行挖掘有可能获得有价值的医学假设。基于文献的知识发现 (Literature Based Discovery, LBD) 是一种通过分析文献, 挖掘潜在知识关联的

方法, 广泛应用于生物医学领域文献挖掘工作。本文总结和梳理 LBD 常用工具、常见方法以及在生物医学领域的应用情况, 指出目前 LBD 研究存在的问题并提出建议。

2 LBD 概述

2.1 定义

1986 年美国芝加哥大学医学教授 Swanson D R 在研究雷诺氏综合症和鱼油时发现两组不存在引文关系的文献之间存在未被注意到的相互联系^[2]。经过深入分析后 Swanson D R 做出雷诺氏综合症和鱼油存在潜在逻辑关系的假设, 因此获得美国情报科学与技术学会最高成就奖^[3]。此后这种通过分析文献发现有价值的隐含关联知识的方法称为 "Literature

〔收稿日期〕 2020-12-10

〔作者简介〕 郎宇翔, 硕士研究生, 发表论文 1 篇; 通讯作者: 杨艳萍, 博士, 副研究员。

Based Discovery"^[4]，在国内经常被称为“非相关文献知识发现”^[3]。近年来大多数 LBD 相关研究已不再强调使用没有引文关系的非相关文献集挖掘隐含关系，因此本文将" Literature Based Discovery" 按照其英文原意翻译为“基于文献的知识发现”。

2.2 ABC 模型

Swanson D R 将文献隐含关系挖掘的方法概括为 ABC 模型^[5]，这也是目前最常见的 LBD 模型。其基本假设是：如果 AB 之间有直接关联，BC 之间有直接关联，那么 AC 之间就被认为有隐含关联。ABC 模型分成两个步骤^[5]：第 1 步，以概念 A 为原点、通过中介术语 B 搜索关联概念 C；第 2 步，挖掘概念 A 和概念 C 之间的关联。国外学者^[6]在 ABC 模型基础上提出开放式和封闭式 ABC 模型。开放式 ABC 模型用于产生关联假设的模型，描述以概念 A 为原点在文件集中发现潜在“A-B-C”关联的过程。封闭式 ABC 模型描述验证假设过程，以概念 A 和概念 C 为起点验证文献集中是否存在同时关联二者的概念 B。ABC 模型是 LBD 研究领域最常见的实现逻辑，但存在较明显缺陷。例如在开放式 ABC 模型中过多中介概念 B 会造成候选目标概念 C 的指数级增长^[7]，从而加大对真正有意义关联的挖掘难度。

2.3 其他 LBD 实现方法

为克服传统 ABC 模型缺陷国内外学者通过使用新工具、引入新方法丰富和扩展 LBD 实现途径。例如 Weeber M、Klein H 和 Aronson A R 等^[8]使用统一医学语言系统（Unified Medical Language System, UMLS）术语代替文献中的单词以表示概念，通过识别文献中非医学概念词实现语义过滤，从而解决候选概念词指数爆炸问题；Wilkowski B、Fiszman M 和 Miller C M 等^[9]扩展 ABC 模型，使用概念链条代替 ABC 模型单一中介概念 B，将 A 和 C 通过一系列中介概念连接起来，使 LBD 方法能够挖掘出更复杂的隐含关联。

3 LBD 研究中的关键问题

3.1 数据获取

3.1.1 文献数据源 主要指专业领域数据库（如 PubMed）、综合商业文献数据平台（如 Web of Science）以及一些公开网站信息等，其中包含文献类型主要有研究论文、工作报告、医疗记录、网站评论以及公司招股说明书等，见表 1。

表 1 文献数据源

文献类型	数据源	数据类型	应用领域	代表文献
研究论文	PubMed/Medline	全文	生物医学	Yetisgen – Yildiz M 和 Pratt W ^[10]
		关键词	生物医学	Jha K、Xun G 和 Wang Y 等 ^[11]
		标题	生物医学	Swanson D R 和 Smalheiser N R ^[12]
		摘要	生物医学	Seki K 和 Mostafa J ^[13]
		标题和摘要	生物医学	Cameron D、Kavuluru R 和 Rindflesch T C 等 ^[14]
		标题、摘要和引文关系	生物医学	Workman T E、Fiszman M 和 Rindflesch T C 等 ^[15]
		全纪录	生物医学	Lindsay R K 和 Gordon M D ^[16]
	SCI	摘要	水净化技术	Kostoff R N、Solka J L 和 Rushenberg R L 等 ^[17]
	SCI/SSCI	标题、摘要和引文关系	生物医学	Kostoff R N ^[18]
		标题、摘要和引文关系	科技与社会	Ittipanuvat V、Fujita K 和 Sakata I 等 ^[19]
工作报告	Elsevier Science Direct	全文	生物医学	Vicente J M ^[20]
	WOS INSPEC 数据库	主题词	航天	曹志杰和冷伏海 ^[21]
	911 委员会报告	全文	政治	Jha K 和 Jin W ^[22]
	医疗记录	临床医疗记录	生物医学	Hanauer D A、Saeed M 和 Zheng K 等 ^[23]
		电子医疗记录	生物医学	Malhotra A、Guendel M 和 Rajput A M 等 ^[24]
网站评论	WebMD	全文	生物医学	Rastegar M 和 Prasad R ^[25]
上市公司招股说明书	-	章节标题	经济与管理	赵一鸣、程斌和王显斌 ^[26]

论文是 LBD 重要挖掘对象，由于工具和资源较为丰富，多数基于论文的 LBD 研究属于生物医学领域，研究内容涉及“药物－疾病”“基因－疾病”“药物－不良反应”等关系挖掘。有研究人员挖掘其他领域论文中的隐含知识，如 Kostoff R N、Solka J L 和 Rushenberg R L 等^[17]检索水净化领域论文，识别文献中的概念、技术组件以发现能提升水净化技术的潜在关联；曹志杰和冷伏海^[19]将 IEE 的 INSPEC 作为数据源进行航天领域潜在关联挖掘，成功识别出一个可能被人为屏蔽掉的等离子体与隐身材料关联。医疗记录是论文之外受关注最多的文献类型，其包含患者基本信息、实验室指标、病程、医嘱记录等多项有效信息，具有较高利用价值^[27]。Shang N^[28]将电子医疗记录结合研究论文来检测药物潜在不良反应。该研究首先从电子医疗记录中检测有关药物不良反应的报告，随后从大量研究论文中挖掘相关信息评估药物不良反应置信度。Hanauer D A、Saeed M 和 Zheng K 等^[23]提出存在于临床医疗记录而在 Medline 记录中缺失的疾病之间具有关联的假设，分别统计临床医疗记录与 Medline 记录中的国际疾病分类号（International Classification of Disease V9，ICD－9）的共现关系，用卡方检验为关联关系进行排序，通过对比二者重叠部分验证假设。随着 LBD 研究的逐步成熟以及自然语言处理等技术进步，工作报告、招股说明书、网站评论等类型文献数据成为 LBD 研究对象，赵一鸣、程斌和王显斌^[26]将上市公司招股说明书作为数据源，抽取其中用于描述风险的主题并挖掘风险间的关联；Jha K 和 Jin W^[22]使用基于图的文本挖掘模型挖掘美国 911 委员会报告并实现 75% 的召回率。

3.1.2 基于文献内容的数据源 经过 2 次加工的增值数据平台，如 SemMedDB 数据库和 MRCOC（MEDLINE Co－occurrences）。其不保存原始文本内容，而是保存从文献中抽取出的概念实体和实体之间的关系。使用该数据源无需在识别和抽取文本概念方面耗费过多精力，有助于简化挖掘隐含关系的过程。SemMedDB 是关系型数据库，保存从 PubMed 文献标题和摘要中获取的概念词。由于 SemMedDB 中的语义谓词与主语、宾语词均来源于 UMLS 标引体系，使得 SemMedDB 能够以标准格式存储不同来源数据^[29]。Hristovski D、Kastrin A 和 Dinevskr D 等^[30]将 SemMedDB 存储的生物医学概念和语义谓词等数据转换成图数据结构，使用 Neo4j 存储图结构并挖掘潜在关联。MRCOC 汇总 Medline 数据平台主题词共现信息，帮助研究人员从 MRCOC 文件直接获取某一篇文章中的主题词共现信息。Spiro A、Fernandez G J 和 Yanover C^[31]使用 MRCOC 中的医学主题词表（Medical Subject Headings，MeSH）关键词共现数据和副作用资源数据库（Side Effect Resource，SIDER）中的药物和不良反应关联数据对隐含药物与副作用关系进行研究。

3.2 概念节点和关系抽取与表示

3.2.1 概述 解决数据获取问题后，从数据中获取感兴趣的概念以及概念之间的关系是 LBD 研究面临的主要问题。大多数现有研究都是利用自然语言处理工具抽取文本中的概念和谓词并使用特定领域叙词表统一表述文本中的概念和谓词。部分成熟工具可以用于从文本中抽取概念和关系，见表 2。

表 2 概念抽取和表示工具

工具	作用	代表文献
MeSH	用于获取概念的标准表述方式和概念之间的层次关系	Huang W、Nakamori Y 和 Wang S Y 等 ^[32]
UMLS	用于识别 MeSH 属于的语义类别	Hu X H、Li G R 和 Yoo I 等 ^[33]
MetaMap	用于识别文本中的 UMLS 概念	Ijaz A Z、Song M 和 Lee D ^[34]
SemRep	生物医学领域自然语言处理工具，用于从文本中抽取出语义谓词	Cairelli M J、Fizsman M 和 Zhang H 等 ^[35]
Stanford Parser	自然语言处理工具，用于对文本进行依存句法分析	Yang H T、Ju J H 和 Wong Y T 等 ^[36]
GloVe	词向量学习工具，用于将生物医学概念表示为实向量	Patrick M T、Raja K 和 Miller K 等 ^[37]

续表 2

Stanford CoreNLP	自然语言处理工具, 用于文本中单词的词干化和词性标注	Song M、Heo G E 和 Ding Y ^[38]
LingPipe	自然语言处理工具, 用于命名体识别	Song M、Heo G E 和 Ding Y ^[38]
ABNER	分子生物学文本分析工具, 用于从文本中识别基因、蛋白质等命名实体	Zhou E、Hui N 和 Shu M 等 ^[39]
Entrez Gene Database	基因数据库, 用于构造从文本中识别基因的命名体识别字典	Kim Y H、Beak S H 和 Charidimou A 等 ^[40]
Therapeutic Target Database	包含靶点、疾病和药物信息的数据库, 用于构造命名体识别字典	Yang H T、Ju J H 和 Wong Y T 等 ^[36]
Kyoto Encyclopedia of Genes and Genomes	用于获取和疾病相关基因	Kim Y H、Beak S H 和 Charidimou A 等 ^[40]
Gene Ontology	基因本体, 用于获取基因功能	Seki K 和 Mostafa J ^[41]

3.2.2 MeSH、UMLS 与 MetaMap 这是生物医学领域 LBD 研究最常用的概念抽取和表示工具, 由美国国家医学图书馆 (National Library of Medicine, NLM) 开发, 专门用于规范化表示生物医学领域概念和实体。其中 MeSH 是被用于标引 PubMed 文献的叙词表, 能规范、唯一地表征生物医学词汇。MeSH 将主题词按照其所属类别组织成树状分类体系, 两个主题词在该分类体系中距离越近则概念越相近。Huang W、Nakamori Y 和 Wang S Y 等^[32]利用 MeSH 对文本中的生物医学概念进行标引, 通过将某个词替换为该词在 MeSH 树状分类系统中的兄弟概念来产生可能的新假设。UMLS 包含叙词表、语义网络、专家词典和一系列工具集^[42]。MetaMap 为 UMLS 中的工具之一, 可识别文本中的生物医学概念并利用 UMLS 语言体系进行标引。SemRep 是 NLM 开发的一款基于 UMLS 的自然语言处理工具, 可以从生物医学文本中抽取出语义谓词^[43], 即“主语-谓语-宾语”三元组, 使用 UMLS 词汇体系进行标引。生物医学命名体识别器 (A Biomedical Named Entity Recognizer, ABNER) 是一个专用于生物医学领域的命名体识别工具, 可识别文本中有关基因、蛋白质等词汇。由于 ABNER 不具有规范化表示概念的功能, 研究人员一般将其与 MeSH、UMLS 等具有叙词表功能的工具配合使用。

3.2.3 专业数据库 Entrez 基因数据库、KEGG 数据库、TTD 数据库以及基因本体等专业数据库可

以帮助识别生物医学文本中的专业概念。其中 Entrez 是分子生物学基因数据库, 保存有核苷酸、蛋白质名称和序列数据。KEGG 是系统分析基因功能知识库, 可将基因组信息与更高阶功能信息连接起来, 保存所有已完全测序的基因组和部分基因组最新基因功能^[44]。Kim Y H、Beak S H 和 Charidimou A 等^[40]使用 Entrez 基因数据库构造命名体识别的字典抽取文本中的基因, 同时利用 KEGG 数据库获取与阿尔兹海默症、帕金森症相关的所有基因, 以对这两种病症相关的潜在概念和关系进行挖掘。TTD 数据库是提供已知治疗性蛋白和核酸靶标、靶向性疾病、通路信息以及针对每个靶标相应药物信息的数据库。Yang H T、Ju J H 和 Wong Y T 等^[36]使用 TTD 数据库构造命名体识别字典以识别文本中的药物、疾病与靶标概念。基因本体可视为有关基因的知识库, 从分子功能、细胞组分、生物过程 3 个角度对基因进行描述。Seki K 和 Mostafa J^[41]在研究中使用基因本体和基因关联数据库获取基因功能以构造基因到疾病的推理网络。

3.2.4 其他抽取工具 除上述几种专门用于处理生物医学领域文本的工具和资源外, 部分研究人员还使用 Stanford Parser、Stanford CoreNLP、LingPipe、GloVe 等无领域限制工具进行关系抽取。其中前 3 种工具可用于识别文本语法结构和概念抽取, GloVe 可以将概念转化为高维实向量, 便于量化表示概念之间的语义关系。

3.3 关系度量与挖掘

3.3.1 概述 LBD 核心任务是从文献中识别概念关联并筛选出有价值、新颖的关联, 实现此任务需要解决关系度量与挖掘问题。研究方法大致经历了从基于词共现、语义关系网络到机器学习的演变。

3.3.2 基于词共现的方法 LBD 最早的实现方法, 通过寻找与初始概念 A 共现的概念 B 进行知识发现。该方法通常使用词频、平均词频、相对文献频率^[45]、平均文献频率^[19]等频繁性度量指标对被挖掘出的关联进行排序。1996 年 Gordon M D 和 Lindsay R K^[46]使用 ABC 模型通过统计 B、C 概念词频率、相对频率等指标将最可能的中介概念 B 和目标概念 C 筛选出来, 成功复制 Swanson D R 在 1986 年的发现, 实现 LBD 过程在一定程度上自动进行; Huang W、Nakamori Y 和 Wang S Y 等^[32]将已经在市场营销方面较为成熟的关联规则挖掘方法应用于 LBD, 使用关系支持度、置信度量化关联的强弱, 筛选出 Medline 记录中重要概念共现关系。

3.3.3 基于语义关系网络的方法 基本思路是识别并抽取文本中属于特定语义角色的概念和概念间的关联并抽象成网络图。该方法通常使用图论方面知识过滤冗余关联、识别关键路径。Hu Z Y、Zeng R Q 和 Qin X C 等^[47]使用 SemMedDB 获取诱导多能干细胞领域文献中的概念并依据概念之间的关系将概念组织成“主-谓-宾”关系网络, 使用子图探测算法挖掘网络中有价值的潜在关联。Vlietstra W J、Zielman R 和 Van Dongen R M 等^[48]在构造完成概念之间的关系网络后, 通过过滤图中与“偏头痛”节点没有直接关联的节点实现过滤冗余链接目的。Wilkowski B、Fiszman M 和 Miller C M 等^[9]在概念网络中选定初始概念和终止概念后使用复合中心度为两个概念节点之间的路径排序, 选出排名前 20 的路径作为 LBD 结果。

3.3.4 机器学习与深度学习相关方法 由于机器学习相关方法需要大量用于模型训练的数据, 因此相关研究重点在于概念之间链接的数字化表示。Kastrin A、Rindfleisch T C 和 Hristovski D^[49]假设相似节点在未来出现关联的可能性更大。该研究将文

献中的概念表示成网络图, 通过共近邻、Jaccard 相似系数、Adamic / Adar、优先链接 (Preferential Attachment) 等指标衡量概念节点之间相似性, 使用有监督、无监督机器学习模型依据上述指标预测两个节点在未来是否会出现关联。Zhao D、Wang J 和 Sang S 等^[50]使用路径排序算法 (Path Ranking Algorithm) 将文献中的“药物-靶点-疾病”路径转化为特征向量以便卷积神经网络训练, 预测出 Typhex 可用于治疗沙门氏菌感染。Patrick M T、Raja K 和 Miller K 等^[37]使用词嵌入技术将文献中的概念转换成高维实向量, 以实现数据在保留原有语义信息基础上可被计算机计算, 预测出治疗银屑病的潜在药物。虽然机器学习是较为新颖的关系挖掘方法, 但在初期数据处理时仍然离不开基于词共现、语义关系网络的传统方法。

3.4 结果评估

3.4.1 概述 一个新的 LBD 系统或者实现方法被提出后研究人员需要验证其是否能从文献中发现隐含关联。目前对预测结果进行评估的方法主要包括重复已有发现、时间切片、基于事实的评估和基于专家的评估 4 种。

3.4.2 重复已有发现 早期 LBD 结果评估常用方法, 一般通过重复 Swanson D R 关于雷诺氏综合症与鱼油、镁与偏头痛的发现检验 LBD 系统能力。该评估方案思路简单, 但不能较好地评估 LBD 系统在其他问题上的表现, 可能造成 LBD 系统只能用于挖掘雷诺氏综合症与鱼油之间的关系。

3.4.3 时间切片法 该方法依据文献是否发表于 T 时刻之前将文献划分为两类, 使用 LBD 系统在 T 时刻前发表的文献中挖掘关联并使用 T 时刻后发表的文献判断上述关联是否被研究验证。若挖掘出的隐含关联被 T 时刻后的研究证实则证明 LBD 系统有效。

3.4.4 基于事实的评估方法 该方法通过判断被 LBD 发现的关联是否已被相关研究证明而实现, 需要事实数据库的辅助, 这些数据库中保存已被研究验证过的已知基因和疾病之间的关系, 见表 3。如 Seki K 和 Mostafa J^[41]使用基因关联数据库中的已

知关系作为金标准（Golden Standard）评估其所提出方法的有效性。Shang N、Xu H 和 Rindflesch T C 等^[51]使用 SIDER 数据库获取目前已被证实的药物和不良反应关系。

表 3 可用于 LBD 的生物医学领域事实数据库

资源	简介	代表文献
Comparative Toxicogenomics Database	存储“基因－蛋白质－药物”关系数据库	Rastegar M、Elayavilli R K 和 Li D 等 ^[52]
Genetic Association Database	存储目前已被验证的“疾病－基因”关系的数据库	Seki K 和 Mostafa J ^[41]
Therapeutic Target Database	存储药物－靶点－疾病之间的关系的数据库	Zhao D、Wang J 和 Sang S 等 ^[50]
Side Effect Resource 2	存储目前已知的“药物－不良反应”关系	Shang N、Xu H 和 Rindflesch T C 等 ^[51]

3.4.5 基于专家的评估方法 将 LBD 系统发现交给领域专家审核，该方法充分利用人的知识，能评估 LBD 系统发现新颖关联的能力。但是其客观性较时间切片法和基于事实的评估方法较低且实现成本较高。

4 LBD 在生物医学领域的应用

4.1 药物潜在关联挖掘

主要包含药物发现、再利用与不良反应挖掘。药物发现是识别潜在新药的过程，相比于其他药物发现方法 LBD 成本更低也更加安全。Swanson D R 利用非相关文献发现雷诺氏综合征和鱼油之间的隐含关联的过程，就是 LBD 在药物发现的最早应用。随着 LBD 的逐步成熟许多新技术被应用于 LBD 药物发现研究中。如 Sang S、Yang Z 和 Wang L 等^[53]从 PubMed 中抽取关系构造知识图谱并设计机器学习模型，从知识图谱中发现潜在药物。药物开发的高成本、高风险性推动了将已被批准的药物应用于其他疾病治疗的尝试，即药物再利用。Rastegar M、Elayavilli R K 和 Li D 等^[54]提出新的 LBD 候选结果排序方法并将其应用于药物再利用研究中。Spiro A、Fernandez G J 和 Yanover C^[31]基于文献内关键词的共现信息预测药物不良反应。

4.2 其他类型生物医学概念间关联挖掘

LBD 还可用于挖掘任意两个生物医学概念之间的关系。例如 Zhou E、Hui N 和 Shu M 等^[39]通过统计 PubMed 论文摘要中 miRNA 与乳腺癌 p53 基因之间的共现频率挖掘两者间潜在关系；Cairelli M J、

Miller C M 和 Fiszman M 等^[55]开展基于语义 Medline 的 LBD 研究，发现肥胖症患者在重症监护室内存活率更高的原因；Hristovski D、Peterlin B 和 Dzeroski S^[56]使用其开发的 LBD 系统发现色素失禁症的潜在致病基因。

5 LBD 研究问题与建议

5.1 数据来源相对单一

各种文献数据库的完善为大量文献数据获取提供可能。但多数 LBD 研究所使用文献类型仍局限于科研论文，对专利、社交媒体文本挖掘相对较少；从文献语言来看，多为英文文本。针对这一问题，未来研究人员应学习和利用自然语言处理领域成熟的开源工具，思考如何为不同类型、语言的文本定制不同的信息抽取与关联挖掘方法。

5.2 缺乏综合性高质量叙词表

美国国家医学图书馆开发的生物医学领域大规模规范化叙词表和自然语言处理工具可以帮助研究人员解决文本中概念和关系抽取问题。拥有这些高质量工具和资源支持是 LBD 被广泛应用于生物医学领域的重要原因。然而这些叙词表大多只适用于英文科研论文，其他语言、类型的文献依然缺乏相应高质量叙词表，因此应重视相关领域叙词表的开发与电子化工作。

5.3 现有方法仍存在缺陷

基于既往研究成果目前已形成多种可行方案解决关联度量与挖掘问题，但各方法均存在固有缺

陷。例如基于词共现的方法未考虑概念之间语义关系；基于语义关系网络的方法需要大量叙词表等资源提供概念语义信息；基于机器学习的方法需要大量标注资源进行训练且其预测的潜在关系可解释性欠佳。因此未来研究需要总结已有方法并尝试融合，形成新方法以弥补已有方法缺陷；借鉴其他领域（如知识图谱、深度学习）最新成果，探索将其应用于 LBD 的可能性。

5.4 评估实例不足

对 LBD 系统预测结果的评估关系到所开发 LBD 系统效果。目前有 4 种常见评估方法，但均存在缺陷：重复已有发现的评估方法只能评价 LBD 系统针对特定问题的性能；基于专家的评估方法受专家主观影响且成本较高；时间切片评估法存在评估结果不一定可靠的问题。文献知识发现的提出者之一 Smalheiser N R 认为开发相当数量的标准实例是解决 LBD 评估问题的关键^[57]。

6 未来生物医学领域 LBD 研究展望

6.1 临床病例的潜在关联挖掘

我国是人口大国，公立医院每年诊疗人次超过 20 亿，我国医疗机构经过长期历史积累已拥有大量电子病历（Electronic Medical Records, EMR）数据。EMR 数据具有采集成本低和数据实时等优势且文本中包含患者病史、家族史、症状以及医生根据症状、理化指标等基础数据做出的诊断等信息^[58]。EMR 这类海量、高价值的数据较适合使用 LBD 相关方法进行潜在关联挖掘。

6.2 中医典籍的潜在关联挖掘

屠呦呦认为应当努力发掘中国医药学资源，其凭借发现青蒿素获得诺贝尔医学奖证明我国中医典籍具有较高研究价值。目前中国中医科学院中医药信息研究所建立的中医药学语言系统（Traditional Chinese Medicine Language System, TCMLS）已初具规模，如何有效利用是该系统目前面临的首要问题^[59]。参考 UMLS 等工具对基于生物医学领域英文

论文的 LBD 研究的促进作用，未来可以利用中医领域相关资源与工具、使用 LBD 相关方法挖掘中医典籍中的潜在关联。

6.3 LBD 在潜在药物发现方面的研究

药物研发具有成本高、周期长的特点。为应对这一问题可以尝试从已有药物中发掘新功效并应用于不同疾病治疗。除将知识图谱作为挖掘潜在药物知识库外，LBD 也可用于潜在药物发现。能够利用最新研究成果挖掘潜在药物，其时效性更强，未来其在潜在药物发现中可能发挥重要作用。

参考文献

- 1 National Library of Medicine. PubMed [EB/OL]. [2020 - 11 - 17]. <https://pubmed.ncbi.nlm.nih.gov>.
- 2 Swanson D R. Fish Oil, Raynaud's Syndrome, and Undiscovered Public Knowledge [J]. *Perspectives in Biology Medicine*, 1986, 30 (1): 7 - 18.
- 3 马明, 武夷山. Don R. Swanson 的情报学学术成就的方法论意义与启示 [J]. *情报学报*, 2003, 22 (3): 259 - 266.
- 4 田瑞强, 姚长青, 潘云涛. 关联文献的知识发现与创新研究进展 [J]. *情报理论与实践*, 2013, 36 (8): 121 - 127.
- 5 Swanson D R, Smalheiser N R. An Interactive System for Finding Complementary Literatures: a stimulus to scientific discovery [J]. *Artificial Intelligence*, 1997, 91 (2): 183 - 203.
- 6 Weeber M, Klein H, Lolkje T W de Jong - van den Berg, et al. Using Concepts in Literature - based Discovery: simulating Swanson's Raynaud - fish oil and migraine - magnesium discoveries [J]. *Journal of the American Society for Information Science and Technology*, 2001, 52 (7): 548 - 557.
- 7 Thilakaratne M, Falkner K, Atapattu T. A Systematic Review on Literature - based Discovery Workflow [J]. *PeerJ Computer Science*, 2019, 52 (6): 1 - 34.
- 8 Weeber M, Klein H, Aronson A R, et al. Text - based Discovery in Biomedicine: the architecture of the DAD - system [EB/OL]. [2020 - 11 - 17]. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2243779/pdf/procamiasymp_00003-0938.pdf.
- 9 Wilkowski B, Fisman M, Miller C M, et al. Graph - based Methods for Discovery Browsing with Semantic Predications [EB/OL]. [2020 - 11 - 17]. <http://europepmc.org/back->

- end/ptpmcrender.fcgi? accid = PMC3243228&blobtype = pdf.
- 10 Yetisgen – Yildiz M, Pratt W. Using Statistical and Knowledge – based Approaches for Literature – based Discovery [J]. *Journal of Biomedical Informatics*, 2006, 39 (6): 600 – 611.
 - 11 Jha K, Xun G, Wang Y, et al. Concepts – Bridges: uncovering conceptual bridges based on biomedical concept evolution [C]. London: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & AMP, 2018: 1599 – 1607.
 - 12 Swanson D R, Smalheiser N R. Implicit Text Linkages between Medline Records: using Arrowsmith as an aid to scientific discovery [J]. *Library Trends*, 1999, 48 (1): 48 – 59.
 - 13 Seki K, Mostafa J. Discovering Implicit Associations among Critical Biological Entities [J]. *International Journal of Data Mining and Bioinformatics*, 2009, 3 (2): 105 – 123.
 - 14 Cameron D, Kavuluru R, Rindflesch T C, et al. Context – driven Automatic Subgraph Creation for Literature – based Discovery [J]. *Journal of Biomedical Informatics*, 2015 (54): 141 – 157.
 - 15 Workman T E, Fiszman M, Rindflesch T C, et al. Framing Serendipitous Information – seeking Behavior for Facilitating Literature – based Discovery: a proposed model [J]. *Journal of the Association for Information Science and Technology*, 2014, 65 (3): 501 – 512.
 - 16 Lindsay R K, Gordon M D. Literature – based Discovery by Lexical Statistics [J]. *Journal of the American Society for Information Science*, 1999, 50 (7): 574 – 587.
 - 17 Kostoff R N, Solka J L, Rushenberg R L, et al. Literature – related Discovery (LRD): water purification [J]. *Technological Forecasting and Social Change*, 2008, 75 (2): 256 – 275.
 - 18 Kostoff R N. Literature – related Discovery: common factors for Parkinson’s disease and Crohn’s disease [J]. *Scientometrics*, 2014, 100 (3): 623 – 657.
 - 19 Ittipanuvat V, Fujita K, Sakata I, et al. Finding Linkage between Technology and Social Issue: a literature based discovery approach [J]. *Journal of Engineering and Technology Management*, 2014 (32): 160 – 184.
 - 20 Vicente J M. The Contribution of Syntactic – semantic Approach to the Search for Complementary Literatures for Scientific or Technical Discovery [J]. *Scientometrics*, 2014, 100 (3): 659 – 673.
 - 21 曹志杰, 冷伏海. 非相关文献知识发现方法在航天科技情报研究中的应用分析 [J]. *情报理论与实践*, 2008, 31 (4): 569 – 572.
 - 22 Jha K, Jin W. Mining Hidden Knowledge from the Counterterrorism Dataset Using Graph – based Approach [M]. Cham: Springer International Publishing, 2016: 310 – 317.
 - 23 Hanauer D A, Saeed M, Zheng K, et al. Applying MetaMap to Medline for Identifying Novel Associations in a Large Clinical Dataset: a feasibility analysis [J]. *Journal of the American Medical Informatics Association*, 2014, 21 (5): 925 – 937.
 - 24 Malhotra A, Guendel M, Rajput A M, et al. Knowledge Retrieval from PubMed Abstracts and Electronic Medical Records with the Multiple Sclerosis Ontology [J]. *Plos One*, 2015, 10 (2): e0116718.
 - 25 Rastegar M, Prasad R. Toward a Complete Database of Drug Repurposing Candidates Extracted from Social Media, Biomedical Literature, and Genetic Data [EB/OL]. [2020 – 11 – 17]. <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&number=7349746>.
 - 26 赵一鸣, 程斌, 王显斌. 面向特定领域的文本知识发现研究——上市公司的风险分类体系及关联识别 [J]. *情报杂志*, 2014, 33 (3): 165 – 170.
 - 27 张凤, 俞佳, 吕良忠. 数据挖掘技术在药品不良反应监测领域中的应用 [J]. *中国药房*, 2014, 25 (46): 4381 – 4384.
 - 28 Shang N. Integrating Domain Knowledge to Improve Signal Detection from Electronic Health Records for Pharmacovigilance [D]. Houston: University of Texas Health Science Center, 2014.
 - 29 Kilicoglu H, Shin D, Fiszman M, et al. SemMedDB: a PubMed – scale repository of biomedical semantic predications [J]. *Bioinformatics*, 2012, 28 (23): 3158 – 3160.
 - 30 Hristovski D, Kastrin A, Dinevski D, et al. Constructing a Graph Database for Semantic Literature – based Discovery [J]. *Stud Health Technol Inform*, 2015 (216): 1094.
 - 31 Spiro A, Fernandez G J, Yanover C. Inferring New Relations between Medical Entities Using Literature Curated Term Co – occurrences [J]. *JAMIA open*, 2019, 2 (3): 378 – 385.
 - 32 Huang W, Nakamori Y, Wang S Y, et al. Fuzzy Predicting New Association Rules from Current Scientific Literature [EB/OL]. [2020 – 11 – 17]. <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&number=1336325>.

- 33 Hu X H, Li G R, Yoo I, et al. A Semantic – based Approach for Mining Undiscovered Public Knowledge from Bio-medical Literature [C]. Beijing: 2005 IEEE International Conference on Granular Computing, 2005: 22 – 27.
- 34 Ijaz A Z, Song M, Lee D. MKEM: a multi – level knowledge emergence model for mining undiscovered public knowledge [J]. *Bmc Bioinformatics*, 2010, 11 (3): 7 – 18.
- 35 Cairelli M J, Fiszman M, Zhang H, et al. Networks of Neuroinjury Semantic Predications to Identify Biomarkers for Mild Traumatic Brain Injury [J]. *Journal of Biomedical Semantics*, 2015, 6 (1): 1 – 14.
- 36 Yang H T, Ju J H, Wong Y T, et al. Literature – based Discovery of New Candidates for Drug Repurposing [J]. *Briefings in Bioinformatics*, 2017, 18 (3): 488 – 497.
- 37 Patrick M T, Raja K, Miller K, et al. Drug Repurposing Prediction for Immune – mediated Cutaneous Diseases using a Word – embedding – based Machine Learning Approach [J]. *Journal of Investigative Dermatology*, 2019, 139 (3): 683 – 691.
- 38 Song M, Heo G E, Ding Y. SemPathFinder: semantic path analysis for discovering publicly unknown knowledge [J]. *Journal of Informetrics*, 2015, 9 (4): 686 – 703.
- 39 Zhou E, Hui N, Shu M, et al. Systematic Analysis of the p53 – related MicroRNAs in Breast Cancer Revealing Their Essential Roles in the Cell Cycle [J]. *Oncology Letters*, 2015, 10 (6): 3488 – 3494.
- 40 Kim Y H, Beak S H, Charidimou A, et al. Discovering New Genes in the Pathways of Common Sporadic Neurodegenerative Diseases: a bioinformatics approach [J]. *Journal of Alzheimers Disease*, 2016, 51 (1): 293 – 312.
- 41 Seki K, Mostafa J. Literature – based Discovery by an Enhanced Information Retrieval Model [C]. Sendai: Proceedings of the 10th International Conference on Discovery Science, 2007: 185 – 196.
- 42 Campbell K E, Oliver D E, Shortliffe E H. The Unified Medical Language System: toward a collaborative approach for solving terminologic problems [J]. *Journal of the American Medical Informatics Association*, 1998, 5 (1): 12 – 16.
- 43 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*, 2004, 36 (6): 462 – 477.
- 44 Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes [J]. *Nucleic Acids Research*, 2000, 28 (1): 27 – 30.
- 45 Thaicharoen S, Altman T, Gardiner K, et al. Discovering Relational Knowledge from Two Disjoint Sets of Literatures Using Inductive Logic Programming [C]. Nashville: 2009 IEEE Symposium on Computational Intelligence and Data Mining, 2009: 283 – 290.
- 46 Gordon M D, Lindsay R K. Toward Discovery Support Systems: a replication, re – examination, and extension of Swanson's work on literature – based discovery of a connection between Raynaud's and fish oil [J]. *Journal of the American Society for Information Science*, 1996, 47 (2): 116 – 128.
- 47 Hu Z Y, Zeng R Q, Qin X C, et al. A Method of Biomedical Knowledge Discovery by Literature Mining Based on SPO Predications: a case study of induced pluripotent stem cells [M]. Cham: Springer, 2018: 383 – 393.
- 48 Vlietstra W J, Zielman R, Van Dongen R M, et al. Automated Extraction of Potential Migraine Biomarkers Using a Semantic Graph [J]. *Journal of Biomedical Informatics*, 2017, 71 (3): 178 – 189.
- 49 Kastrin A, Rindflesch T C, Hristovski D. Link Prediction on a Network of Co – occurring MeSH Terms: towards literature – based discovery [J]. *Methods of Information in Medicine*, 2016, 55 (4): 340 – 346.
- 50 Zhao D, Wang J, Sang S, et al. Relation Path Feature Embedding Based Convolutional Neural Network Method for Drug Discovery [J]. *Bmc Medical Informatics and Decision Making*, 2019, 19 (S2): 59.
- 51 Shang N, Xu H, Rindflesch T C, et al. Identifying Plausible Adverse Drug Reactions Using Knowledge Extracted from the Literature [J]. *Journal of Biomedical Informatics*, 2014 (52): 293 – 310.
- 52 Rastegar M, Elayavilli R K, Li D, et al. Assessing the Need of Discourse – level Analysis in Identifying Evidence of Drug – disease Relations in Scientific Literature [J]. *Studies in Health Technology and Informatics*, 2015, 216 (1): 539 – 543.
- 53 Sang S, Yang Z, Wang L, et al. SemaTyP: a knowledge graph based literature mining method for drug discovery [J]. *Bmc Bioinformatics*, 2018, 19 (1): 193.

(下转第 47 页)

- 10 李森, 王玲, 贾晓君, 等. 不同时期临床护士使用移动护理信息系统满意度的比较分析 [J]. 中国护理管理, 2014 (2): 150-153.
- 11 蒋和宏. 重庆市使用新型毒品的男男性行为者艾滋病感染状况及影响因素研究 [D]. 重庆: 重庆医科大学, 2017.
- 12 胡慧航. 社会资本对个人知识组织化的影响——基于组织学习视角 [D]. 杭州: 浙江工业大学, 2011.
- 13 魏莎. 基于技术接受模型研究护士接受移动护理信息系统的影响因素 [D]. 重庆: 重庆医科大学, 2016.
- 14 刘鲁川, 孙凯. SaaS 外包服务用户满意度与持续使用的实证研究 [J]. 信息资源管理学报, 2012 (1): 26-32.
- 15 潘霞. 循证护理模式在临床护理教学中的初探 [J]. 临床医药文献电子志, 2019, 6 (54): 179.
- 16 陈欣, 周晶, 罗力. 国内 MOOC 用户接受影响因素研究 [J]. 计算机教育, 2016 (4): 63-67.
- 17 刘小青, 沈敏, 王金玉, 等. 护理信息小组的建立与实践成效 [J]. 护士进修杂志, 2019, 34 (11): 995-998.
- 18 周丹, 尹安春. 人工神经网络在护理领域的应用研究进展 [J]. 护理学杂志, 2020, 35 (3): 94-97.
- 19 张晓. 长期护理保险智能信息系统建设探讨——基于基本医疗保险的启示 [J]. 信息系统工程, 2019 (2): 36-37.
- 20 王晓慧, 向运华. 老年智慧照护服务体系探究 [J]. 学习与实践, 2019 (5): 88-97.
- 21 严爱龙. 智能化临床护理管理信息系统建设 [J]. 自动化与仪器仪表, 2018 (6): 155-157.
- 22 祁少镔. 信息技术助力护理质控 [J]. 中国新技术新产品, 2019 (17): 34-35.
- 23 常光耀, 黄彩红, 王奕. 基于智能终端的移动护理信息系统应用价值研究 [J]. 中国实用护理杂志, 2016, 32 (24): 1903-1905.
- 24 勾成俊, 于士坤, 彭逢安, 等. 基于物联网的智能护理信息系统研制与护士满意度调查 [J]. 分子影像学杂志, 2015, 38 (2): 158-161.
- 25 张霞, 华皎, 陈洁. 基于 WIFI 和 RFID 技术的智能终端 SDA 在实现患者安全管理目标中的应用 [J]. 当代医学, 2019, 25 (2): 182-184.
- 26 赵延红, 梁宗强, 田卫峰, 等. 临床护理智能交互系统的建设与实践 [J]. 中国数字医学, 2016, 11 (4): 10-12.
- 27 赵浩宇, 周琳, 吴昊. 智能输液监测系统设计与应用 [J]. 医学信息学杂志, 2018, 39 (9): 32-36.
- 28 何蕾, 于卫华, 黄竞竞, 等. 智能一体化护理信息平台的构建与应用 [J]. 齐鲁护理杂志, 2016, 22 (23): 94-95.
- 29 李红莉, 杨雅. 基于信息化系统临床护理智能显示平台的构建及应用 [J]. 护理学报, 2018 (14): 29-33.
- 30 田雨同. 欧美发达国家护理信息系统应用现状 [J]. 医学信息学杂志, 2019, 40 (11): 13-18.
- 31 赵婷婷, 潘璐意, 万光明, 等. 智能护理信息系统在胸痛中心急诊抢救中的应用 [J]. 中国医药导报, 2020, 17 (16): 167-170.

(上接第 41 页)

- 54 Rastegar M, Elayavilli R K, Li D, et al. A New Method for Prioritizing Drug Repositioning Candidates Extracted by Literature-based Discovery [EB/OL]. [2020-11-17]. <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7359766>.
- 55 Cairelli M J, Miller C M, Fiszman M, et al. Semantic MEDLINE for Discovery Browsing: using semantic predication and the literature-based discovery paradigm to elucidate a mechanism for the obesity paradox [J]. AMIA Annual Symposium Proceedings, 2013 (6): 164-173.
- 56 Hristovski D, Peterlin B, Dzeroski S. Literature Based Discovery Support System and Its Application to Disease Gene Identification [M]. Berlin: Springer, 2007: 307-326.
- 57 Smalheiser N R. Literature-based Discovery: beyond the ABCs [J]. Journal of the American Society for Information Science and Technology, 2012, 63 (2): 218-224.
- 58 阮彤, 高炬, 冯东雷, 等. 基于电子病历的临床医疗大数据挖掘流程与方法 [J]. 大数据, 2017, 3 (5): 83-98.
- 59 贾李蓉, 朱玲, 董燕, 等. 中医药学语言系统评价体系的研究与建立 [J]. 中国数字医学, 2012, 7 (10): 13-16.