面向知识发现的生物医学文献信息检索与可视化设计

张 莉 闵 波 杨 帆 张云宏 杜 冰 许文娟

（新疆军区总医院 乌鲁木齐 830000）

〔摘要〕 探讨生物医学文献信息检索过程与可视化设计对用户知识发现的影响与作用，从输入输出、搜索过程、信息定位几方面介绍文献信息检索过程交互流模型，分析文献信息可视化设计的方法，包括颜色与字体、统计图表、分类列表与标签、层次结构与网络图等方面。

〔关键词〕 信息检索；文献挖掘；信息可视化；知识发现

〔中图分类号〕R－056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2017.12.014

Knowledge Discovery－oriented Biomedical Literature Information Retrieval and Visualization Design

ZHANG Li, MIN Bo, YANG Fan, ZHANG Yun－hong, DU Bing, XU Wen－juan, General Hospital of Xinjiang Military Region, Urumqi 830000, China

〔**Abstract**〕 The paper discusses influence and effect of information retrieval process of biomedical literature and visualization design on user knowledge discovery, from the aspects of input and output, retieval process and information positioning, it introduces the interaction model of literature information retrieval process and analyzes approach of literature information visualization design, including color and font, statistical chart, list of categories and labels, layer structure and network pictures, etc.

〔**Keywords**〕 Information retrieval; Literature mining; Information visualization; Knowledge discovery

1 引言

科技文献检索作为科研工作的重要内容，文献数据库服务极大提高检索效率的同时，数量的激增也对知识发现提出新要求，如 PubMed 现有的文献量已接近 3 000 万，面对如此海量的文献数据，传统的文献检索已不足以满足科研人员高效获取知识的需求。为更好地利用丰富的生物医学文献数据，国内外的科研工作者针对不同的研究课题与对象，开始关注从不同角度实现信息提取与知识发现。本文从信息检索与可视化的角度，如何更好地辅助科研人员提高知识发现的效率，探讨生物医学文献信息检索过程与可视化设计对用户知识发现的影响与作用，分析常用的文献信息检索结果的信息可视化设计与呈现方法。

2 文献信息检索

2.1 概述

生物医学文献信息检索是科研工作者根据学习和工作的需求查找与某一特定主题相关的文献资源的过程，对输入的主题关键词具有很强的依赖性，当返回的候选结果数量巨大时，需进一步人工筛选才能准确找到所需文献，见图 1。文献信息检索的

〔收稿日期〕 2017－08－29

〔作者简介〕 张莉，副主任医师；通讯作者：闵波。

过程实际是一个不断交互的启发式知识发现过程，主要包括输入输出、搜索过程与信息定位的 3 部分组成，其中每一部分对文献知识发现都具有非常重要的作用，见图 1。

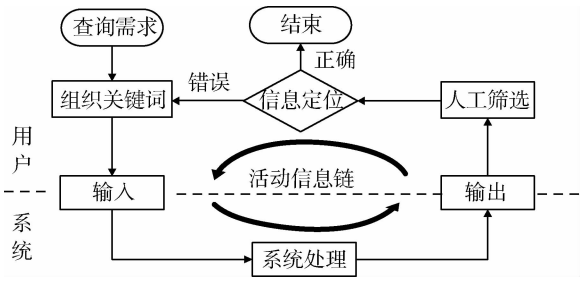

图 1 信息检索交互流模型

2.2 输入输出

输入表达用户的查询需求，通过组装关键词给出检索条件，不同检索平台将输入理解为不同的查询限定条件对象，采用不同的检索策略进行处理[1]。如 iHOP 将单独的基因、蛋白、化合物、疾病等主题词，作为一个生物实体概念去检索相关的文献和信息，可以找到概念间语义相关的目标文献。输出既依赖于输入关键词，也依赖于平台的检索模型与索引算法，输出的信息经过进一步的加工总结，加强用户对结果的认识与理解，达到提升知识发现的效率的目的，如 GOPubMed 使用本体论的背景知识对大量的查询结果的简要概括，便于用户对查询结果有整体把握与总体视角[2]。

2.3 搜索过程

当用户缺乏完整的检索词项时，要准确找到特定主题相关的信息，需要不断根据搜索反馈修改关键词与筛选候选结果集，因此整个检索交互的过程是否友好直接影响检索效率与用户对前端信息的及时理解。iPubMed[3] 针对客户端到服务器数据传输、网络带宽与查询计算所带来的时间延迟，开发相应的索引结构与查找算法，实现检索操作“即输即现”的动态交互，通过动态响应关键词的修改，即时刷新输出结果，达到快速筛选与定位所需文献的目的。XplorMed[4] 对返回的结果集进行总结，使用户能够专注感兴趣的项，通过抽取与输入关键词、相关词，作为候选的检索词，从关键词到相关词再到结果浏览，整个过程形成一个关键词引导信息链。交互过程中良好的提示与引导、动态交互都能够缩小导航路径，减少查找时间，达到辅助用户快速定位信息的目的。

2.4 信息定位

文献检索服务能够帮助用户查询所需的文献，但是文献中的信息才是用户的最终需求，如果能够对文献的主题与内容进行适当的信息抽取，可以帮助用户快速定位到关注的信息，如段落、句子、图片等。BioText[5] 能够直接检索文献中的图片与注释，通过增加与文本内容相关联的图片信息，可以帮助用户对检索到相关性更强的文献。PIE[6] 利用机器学习的方法进行单词与语法分析，实现蛋白质相互作用信息抽取，输出蛋白相互作用相关的文献，按照蛋白质相互作用的可信度排序。MarkerInfoFinder[7] 能够在文献中自动识别疾病名称和不同类型的遗传或遗传变异标记信息。由于传统文献检索不能有效地处理多项查询，难以直接获得给定的概念间关系的文字证据，特别容易忽略概念间可能隐含着的语义关系，因此 MedEvi[8] 利用匹配的多个关键词共现的位置限制，探索位置相互靠近的词之间隐含的语义关系，这显然对形成新的科学假设具有积极的影响。

3 文献信息可视化

3.1 概述

信息可视化是一种利用计算机支撑的、交互的、对抽象数据的可视表示，来增强人们对抽象信息认知的方法，通过对挖掘的信息进行图形与交互地输出显示，提升用户体验，使用户可以高效地与大量数据互动，更好地理解信息与发现隐含的知识，引导出新的假设[9]。在生物医学文献信息检索过程中，利用信息可视化的方法，根据数据的结构特征、模式及其各种语义关系转换成图形，在有限的输出界面中较好地组织，以直观的图形化方式显示出来，实现对信息的分类、导航、缩小查找范围

与高效交互，使用户能快速地从中发现有效的知识。随着文献的数量增长与深度挖掘的需要，越来越多直接或间接的模式，对信息显示的方式提出新的要求。

3.2 颜色与字体

对于核心内容为文本数据的文献来说，颜色与字体可以较好地被用于呈现差异信息，突出重点信息。目前文献都具有统一的结构特征如标题、摘要、关键词、期刊、日期等，而且不同的特征通常具有不同的信息权重。文献检索任务主要都是基于标题关键词来查找文献，因此文献结构特征可以作为前端界面的可视元素，用于增强用户对结果的信息搜索与定位的体验。最常见的方式是通过不同的字体或颜色编码，对不同特征元素采用不同的样式显示，如对标题与摘要用不同大小的字体显示，对匹配的关键词进行高亮显示。特别是用户快速交互的过程中，突出的信息可以形成对视觉的较好刺激，帮助发现与定位信息，如 iPubMed 服务中的即输即现的模式，高亮显示方式在快速响应的同时很好地帮助用户定位目标结果。

3.3 统计图表

面对用户模糊检索的需求时，对候选集提供一定的统计分析信息，对用户具有引导与提示的作用及分类与导航的作用，如最简单的总数统计功能，有助于用户对是否直接人工筛选还是改进检索策略具有很好的提示作用。GOPubMed 利用条形图来显示某个关键词查询结果中文献数量排名前 10 位的机构与期刊的统计信息，利用线形图显示不同时期的相关文献数量的变化趋势。统计图表的使用不仅描述结果集的整体信息，而且可以有效地组织信息，达到分类与导航的功能，缩小查找信息的范围。前端界面上统计图表表的可视化交互功能，提升链接到所选信息的效率，达到快速浏览的目的。

3.4 分类列表与标签

当文献检索返回的候选结果数量巨大时，将信息按一定特征分类或分组，可有效地缩小信息查找范围，减少人工筛选的时间。检索过程中，根据文献的学科属性特征分类，组织输出的信息，设置类型列表栏目，如按期刊分类、按时间排序或分类等，用户可以按需求从相应类别中进一步筛选。此外，文献检索或抽取出来的信息，根据其不同的特征，赋予一定意义的标签，如标签云的应用，对结果集合中的不同词项、信息单元的显示样式进行视觉加权。标签的使用是用户对信息的一种标记或记录，此方法不仅使得信息可以更好地被找到，而且实现对某些信息进行的总结与概括，使得后续用户可以更好地理解信息。

3.5 层次结构与网络图

领域文献之间存在着复杂的知识相关性，如主题相关性、引用关系与学科分类关系等，这些内在的语义关系不但将文献中存在的信息之间建立某种联系，也对文献信息检索起到了导航与引导的作用。GOPubMed 将结果集与 Gene Ontology 的层次结构整合，通过 Gene Ontology 的层次结构与分类关系，用户能够按类别快速浏览摘要，方便导航与筛选文献。PubNet[10] 是一个用于抽取不同类型关系的文献挖掘平台，基于关键词检索抽取出的关系映射到文献衍生网络中，网络结构特征直观形象地显示出这些不同对象间的关系。iHOP[11] 通过使用基因和蛋白质作为句子或摘要之间的超链接，将文献中的信息转换为一个如同互联网的导航资源，而且基因与蛋白质间的生物学关系可以将文献中文本信息映射到生物实体关系网络。目前，利用网络模型进行文献知识发现研究已成为生物医学信息学领域的热点内容。

4 结语

随着文献数量的海量增长，面向知识发现的文献信息检索是大数据时代新的热点需求，而构建一个完善的生物医学文献知识服务平台，需要计算机、生物学与语言学等多学科专家的共同参与。大量科学家利用文献库与生物信息数据库进行整合，从而完善概念实体的识别，以便挖掘出更加完整可

靠的信息。信息可视化作为一项强实践性的工作，通过将信息进行视觉化呈现，提高知识的可感性与可见性，但不同的可视化方法很难准确评价，只能针对特定的问题与需求采用特定的可视化方式；研究新的高维的多层次的大数据集可视化方法，引入生物学意义的视角与维度，可以帮助发现深层次的知识，促进新科学假设的形成。现阶段文献挖掘研究正处于快速发展时期，许多新的方法与技术在不断地被开发出来，文献检索在不断完善信息查找的精确性的同时，正在向科学假设自动形成的知识服务发展。

参考文献

1 Kim JJ, Rebholz - Schuhmann D. Categorization of Services for Seeking Information in Biomedical Literature: a typology for improvement of practice [J]. Brief Bioinform, 2008, 9 (6): 452 - 465.

2 Doms A, Schroeder M. GoPubMed: exploring PubMed with the Gene Ontology [J]. Nucleic Acids Res, 2005, 33 (Web Server issue): 783 - 786.

3 Wang J, Cetindil I, Ji S, et al. Interactive and Fuzzy Search: a dynamic way to explore MEDLINE [J]. Bioinformatics, 2010, 26 (18): 2321 - 2327.

4 Perez - Iratxeta C, Pérez AJ, Bork P, et al. Update on XplorMed: a web server for exploring scientific literature [J]. Nucleic Acids Res, 2003, 31 (13): 3866 - 3868.

5 Hearst MA, Divoli A, Guturu H, et al. BioText Search Engine: beyond abstract search [J]. Bioinformatics, 2007, 23 (16): 2196 - 2197.

6 Kim S, Kwon D, Shin SY, et al. PIE the Search: searching PubMed literature for protein interaction information [J]. Bioinformatics, 2012, 28 (4): 597 - 598.

7 Xuan W, Wang P, Watson SJ, et al. Medline Search Engine for Finding Genetic Markers with Biological Significance [J]. Bioinformatics, 2007, 23 (18): 2477 - 2484.

8 Kim JJ, Pezik P, Rebholz - Schuhmann D. MedEvi: retrieving textual evidence of relations between biomedical concepts from Medline [J]. Bioinformatics, 2008, 24 (11): 1410 - 1412.

9 王敏，张燕舞，张玢，等. 信息可视化在医学文献分析中的初步应用理论研究 [J]. 医学信息学杂志，2010，31 (2)：40 - 44，49.

10 Douglas SM, Montelione GT, Gerstein M. PubNet: a flexible system for visualizing literature derived networks [J]. Genome Biol, 2005, 6 (9): R80.

11 Hoffmann R, Valencia A. Implementing the iHOP Concept for Navigation of Biomedical Literature, Bioinformatics, 2005, 21 (Suppl 2): 252 - 258.

(上接第 64 页)

3 Chen C. CiteSpace 1I: detecting and visualizing emerging trends and transient patterns in scientific literature [J]. Journal of the American Society for Information Science and Technology, 2006, 57 (3): 359 - 377.

4 Chen C. Searching for Intellectual Turning Points: progressiveknowledge domain visualization [J]. Proceedings of the National Academy of Sciences, 2004, 101 (1): 5303 - 5310.

5 刘雪立. 基于 Web of Science 和 ESI 数据库高被引论文的界定方法 [J]. 中国科技期刊研究，2012，23 (6)：975 - 978.

6 焦宏官. 基于 SCIE 的国际针灸热点及合作团队研究 [R]. 北京：中国中医科学院中医药信息研究所，2013：95.

7 高丽丽，郭义. 近5年SCI源期刊发表有关针灸文章的总结分析 [J]. 湖南中医杂志，2013，29 (4)：144 - 146.

8 王雪苔. 论针灸特色 [J]. 中国针灸，2005，25 (2)：75 - 78.