非相关文献知识发现法在中医研究中的应用

★ 邵运峰 （江西中医学院图书馆 南昌 330006）
★ 翁捷 （江西科技师范学院 南昌 330013）

关键词：非相关文献；软件
中图分类号：G 354.4 文献标识码：A

随着学术研究的领域拓宽、内容深化，科技文献的数量飞速增长。在海量增长的文献面前，研究者个人所能涉猎的文献显得非常有限。即便是在文献存贮、利用完全数字化的今天，研究者对文献内容的处理能力仍显不足。这迫使文献情报工作者和科研工作者开始寻找一种能直接对文献内容进行分析处理的情报学手段。非相关性文献知识发现法就是这样一种全新的文献情报学方法。

1 非相关文献的概念

科学文献间有着错综复杂的联系，有些联系是显性的，是为人们所共知的，通过引文分析等方法可以判断出它们的关联情况：例如文献间引文与被引的互引关系；两篇或多篇文献共同引用一篇文献的共引关系等，我们称这种表面上具有显性关联的文献为相关文献。与相关文献相反，有些文献之间使用引文分析等方法来判别，不存在关联性，但却可能存在着尚不为人们所知的潜在的关联，这类文献我们称为非相关文献。

非相关文献广泛地存在于学科与学科之间，亦存在于同一学科内不同的专业领域之间。当今中医的研究越来越深入，学科不断裂化，专业研究领域越来越专业化，新领域越来越多。某专业领域的文献，在一般情况下并不为另一专业领域的研究人员所熟知，几乎没有人会想到将两类文献放在一起加以研

十年的中医发展战略似乎都因缺乏冷静客观的思索，犯了激进浮躁的毛病。

姚老在早年改进计划中提出了中医发展“三步走”的战略设想，提出“从原有基础上提高一步”的着手步骤，我以为切中时弊、稳妥扎实，今天看来，仍有极大的借鉴意义。

3.1 “三步走”的战略设想 首先运用科学方法从速搜集整理固有学说，使之成为具体的知识系统；进而利用各种自然科学知识，求取实证，说明原理；最后再与西医根本沟通，互相促进，融合统一，共同发掘未知。

中医科学化，首先最急需、也最可行的目标是有计划地系统搜集、整理中医原有的学说，认清已有基础，使之成为更加具体明确的学术体系。这既能使中医整体水平在原有基础上提高一大步，也能为在自然科学上求取证实、解释，提供比较清楚的研究背景和带根本性的研究靶点。而与西医的根本沟通，必待中医原理得到自然科学的解答之后才能彻底实现，否则，面对中医以“气化”一贯到底的错杂学理，一味想与西医作现成对接，只能是拼凑强同，结果不但不能提高，反会使中医原有的治疗水准普遍下降。

3.2 “从原有基础上提高一步”的着手步骤 (1)中医自身改进任务——文献＋临床研究为主：分类搜集原有学说；分类整理原有学说；评定初步教材；调查实验初步教材。

(2)汇同其他学科人才——临床＋实验研究为主：利用自然科学沟通说明；统一发掘尚未解答之实例；确定进步教材。

中医科学化，首先就面临着在新的历史条件下中医传统特色的科学继承问题。我们常感慨，中医临床水平有一代不如一代的趋势。为此，我曾求教于姚老，先生则语重心长告之：“责之原因，似乎是学生没学好，其实背后是老师的原因，而说老师没教好，其实背后又是教材的原因。教材挂一漏万，脱离实际，看起来是编委的责任，而归根到底，乃在于中医整体对原有学说没有真正全面地分类搜集、系统整理。因此，任何主编白发穷年，都不足以反映和代表中医应有的实际水平。加之，实际编写仓促、浮躁，教材又何以能有真正的权威性？”由此话再联想到中医教育一直存在困惑，更感到先生“从原有基础上提高一步”的改进工作，是多么的现实与必要！如今，难道我们中医队伍自身，还不应该以全国一盘棋的思想，群策群力，补救完成这项“欠债”已久而又责无旁贷的历史任务吗？

（收稿日期：2005-02-23）

究,即使将两类文献放在一起,仅仅依靠研究者的脑力劳动也较难发现其中可能存在的各种隐含联系,要充分揭示非相关文献的隐含联系,必须借助知识发现的理论和方法。

2 非相关文献知识发现法在中医研究中的应用前景

中医药有很悠久的历史,古文献很多,近现代使用新技术、新方法对中医药进行研究也产生了大量的文献。但是中医药学科内仍然有大量的问题,甚至是一些基本性的问题得不到合理的、科学的解释,研究的空白比比皆是。新技术、新方法的引入,是中医药学科发展当中相当紧迫的任务。非相关文献知识发现法是一种全新的、独特的情报学方法,对文献的有效使用、文献中隐藏知识的发掘,可以起到较大的作用。将其引入到中医药学科的研究当中,将能对中医药学科文献的有效利用起到巨大的推动作用。

3 非相关文献知识发现法的实现工具——Arrowsmith 系统

Arrowsmith 系统是实现非相关文献知识发现的软件工具,可以免费使用,最新的版本是 3.0,网址是 http://kiwi.uchicago.edu 或 http://arrowsmith.psych.uic.edu。其主要功能是:从两类非相关文献数据库记录的标题、主题词及文摘当中,提取自然语言并加以分析排列,找到能表达两类非相关文献间关联性的概念、词语等,供研究人员参考。

研究人员可以借助 Arrowsmith 系统建立科学假说或者启发思路,或者在一定程度上验证科学假说。Arrowsmith 系统使用了停用词表、语义筛选、词频统计等方法,对非相关文献之间的关联性进行智能的初步筛选,生成一个关联词列表。通过 Arrowsmith 系统,人们对文献资源内在的关联性进行严密分析,得到的结果有很强的针对性和方向性,使科研工作的效率得到提高,避免了一些繁琐、简单的工作。

4 非相关文献知识发现中文工具软件的研制

Arrowsmith 系统作为非相关文献知识发现的有效工具,与很多的大型文献数据库系统实现了紧密的结合,例如 Medline、Biosis、Embase、Scisearch Internet databases 等。虽然随着中医药学科在世界上影响的日益扩大,上述数据库中也都收录有中医药研究的文献,但是,收录的文献量是相当小的,中文研究文献就更少。而且,Arrowsmith 系统是处理英文文献的系统,而处理中文文献需要不同的分词技术和词表系统。因此 Arrowsmith 系统在中医药学科研究中的直接作用,是相当有限的。但是,我们完全可以借鉴和参照 Arrowsmith 系统,研制可处理中文中医药学非相关文献的软件系统。

非相关文献知识发现软件系统本身并不复杂,但是其实现需要依赖几项关键技术:

(1)非相关文献的自动判别:文献间非相关关系的确定主要依赖引文分析法,排除文献间存在的互引、共引等关系。目前,在"中文科技期刊数据库(引文版)"、"中国期刊全文数据库"等数据库中,对文献间的引文关系都有所揭示,为确定非相关文献提供了良好的数据基础。但是,对引文数据进行准确提取,以及设计相应的算法,仍需要做大量的工作。

(2)中医药学中文文本自动分词:Arrowsmith 系统处理非相关文献的第一步,就是要从两个非相关文献题名、主题词、文摘的集合中,抽取自然语言。由于中文文本不具备词组边界,因此从特定文本里提取有效的关键词就较英文资料困难许多。近年来,中文自动分词技术,无论是自动分词的算法方面,还是词表的研制方面,都取得较大进展,技术已经日益成熟。将这些技术应用于非相关文献研究当中还需要做一些软件实现或集成的工作。

(3)停用词表研制:在非相关文献处理过程中,所需要抽取的语词,并不应该是非相关文献题名、主题词、文摘中包含的全部语词。有一些无意义的语词,或者是在特定研究领域内无意义的语词,应该列入停用词表,在抽词的过程加以删除,降低运算的复杂程度,提高结果的准确性。

5 正确认识非相关文献知识发现法

非相关文献知识发现法,是一种全新的情报学方法,在科学研究当中也被证实为一种行之有效的辅助手段,可以为科研工提供有益的提示,提高效率,成为科学研究的一条捷径。但是非相关文献知识发现法,并不能替代传统的情报检索方法,其所揭示的文献间的关联是否真实可靠,仍然需要通过科学实验加以验证。

参考文献

[1]李亚星,刘莉."尿布与啤酒"对医学科研的启示[J].医学与哲学,2004,25(6):55

[2]董风华,兰小筠.基于文献的科学假说[J].医学与哲学,2004,25(6):57

[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

[4]许建阳,马明,工发强.Swanson 的非相关文献知识发现法对医学发展的思考[J].医学与哲学,2003,24(8):21

(收稿日期:2004-12-03)