刍议期刊文献的专题大数据挖掘价值
——以《临床误诊误治》杂志误诊数据论文策划为例

丁 滨 陈晓红†
《临床误诊误治》杂志社,050082,石家庄

摘 要 介绍误诊疾病数据库研究背景及现况,总结《临床误诊误治杂志》数据库论文策划的体会。通过研制误诊疾病数据库,深入挖掘误诊文献数据价值。已收录2004—2013年2.7万篇标准误诊文献,可生成2 400个疾病误诊数据分析。基于数据库产出的误诊大数据,2015年以来已组织编发12篇单病种误诊研究的数据库论文。

关键词 误诊;数据库;大数据;数据库论文

Discussion on the data mining value in the thematic database of periodical literature: the data article planning in journal of *Clinical Misdiagnosis and Mistherapy* as an example // DING Bin, CHEN Xiaohong

Abstract Introducing the background and current situation of the research on the database of misdiagnosis diseases, this paper summarizes the experience in the planning about the journal database of *Clinical Misdiagnosis and Mistherapy*. Through the development of the database of misdiagnosed diseases, the value of the misdiagnosis literature data was deeply excavated. Twenty-seven thousand papers of standard misdiagnosis literature from 2004 to 2013 have been included, and can generate 2400 data analysis items about misdiagnosed diseases. Based on the misdiagnosis data from the database, 12 database articles about single disease misdiagnosis have been published since 2015.

Keywords Misdiagnosis; database; big data; database article

Author's address Magazine office of Clinical Misdiagnosis and Mistherapy, 050082, Shijiazhuang, China

DOI: 10.16811/j.cnki.1001-4314.2016.05.025

未来,医学大数据分析将越来越左右临床医学的指南、共识和多中心研究,并将产出海量各种形式的数据论文。近年,学术期刊如何应对大数据时代的挑战,现智能添加。下一步要继续研究采集器的其他功能,并开发其深度数据挖掘的功能,以进一步提高过刊网刊元数据的提取效率,降低成本。

2.2.2 将元数据导入投稿系统 数据提取完成后可以导出的形式有:1)导出到数据库;2)导出到网站;3)导出到excel 2007;4)导出到excel 2003;5)导出到Txt;6)导出到html。编辑可以根据各自期刊的情况选择相应的导出格式,然后上传到投稿系统的发行中心中。

3 结束语

稿件加工完成后最麻烦、最费事的工作就是网刊元数据的提取和上传,提取过程中经常会出现各种各样的问题。为了方便作者下载和提升期刊的影响力,网刊元数据的提取和上传又是一项必须执行的工作;因此,怎样提高数据提取的效率,是数字出版编辑需要研究的重点内容。编辑人员只需借助八爪鱼软件,根据期刊的情况设置一定的规则,对导出的数据稍做调整,就可以既省时又省力地完成对网刊元的提取和上传,并且规则设定之后可以重复使用,既克服了传统手工标记时工作量大和容易出错的问题,又极大地提高了后期数据提取的效率。

但是,八爪鱼采集器只能根据网址采集网页上存在的数据,对于网页上没有的信息则无法提取,无法实

4 参考文献

[1] 仇玉坤,顾冠华. 期刊被网络数据库"拆零重组"与对策[J]. 编辑之友,2015(1):54
[2] 潘霄. 中文专家元数据提取研究[D]. 昆明:昆明理工大学,2014
[3] 刘华中. 面向PDF文档的论文元数据提取方法研究[D]. 秦皇岛:燕山大学,2012
[4] 崔玉洁. 我国高校数字期刊发展进展与反思:结合西南大学期刊社数字出版实践[J]. 西南农业大学学报(社会科学版),2013,11(9):194
[5] 俞菁,陈波. 互联网+科技期刊出版:构建科技期刊编校平台的设想[J]. 编辑学报,2015,27(6):590
[6] 杨海亮,徐用吉. 利用VB读取方正排版文件提取元数据[J]. 中国科技期刊研究,2015,26(6):612
[7] 张铭,银平,邓志鸿,等. SVM+BiHMM:基于统计方法的元数据提取混合模型[J]. 软件学报,2008,19(2):358
[8] 韩云波,蒋登科. 参考文献国家标准GB/T 7714—2015的修订特色与细则商榷[J]. 西南大学学报(社会科学版),2015,41(6):157

† 通信作者

(2016-03-06 收稿;2016-04-19 修回)

也成为期刊同道关注的热点。有学者提出学术期刊应从数字化到数据化转型,但纵观近年我国科技出版学者对大数据时代科技期刊生存与发展的研究,还在解读大数据、探讨研究学术期刊由数字化向数据化转型的必要性,具体实践方面的总结较少[1-4]。2015 年以来,《临床误诊误治》杂志依托"误诊疾病数据库检索及管理系统",组织策划了一批具备数据库论文主要特征的误诊数据库单病种误诊研究选题,现做一初步总结。

1　大数据时代学术期刊面临的挑战

1.1　数据论文的概念　"大数据"引起了研究模式的革命性变化,数据论文(data paper) 应运而生。数据论文的概念早已有之。早在 2000 年,美国生态学会就在投稿须知中对其进行了阐释:数据论文是一种特殊类型的论文,用于展示大型或丰富的数据集,包括描述数据内容、数据产生背景、数据质量和结构的元数据文件。随着数据共享理念的发展和在线共享方式的普及,Chavan 和 Penev 于 2011 年将其概念发展为:数据论文应该总是与其描述的公开发布的数据集链接。概言之,数据论文是对在线数据集(dataset) 或一组数据集进行描述的元数据文档,遵循一定的数据标准,计算机可读、可检索[5]。数据论文的基本特点是重点关注数据本身,有专业的数据存储平台,有对数据标准的规定和说明,对数据的管理、共享、传播、重用、演绎等权益做明确说明,等等[6]。

1.2　数据出版的内涵　有学者指出,学术期刊大数据出版可以理解为符合大数据思维模式、理念和精神的学术期刊数字出版行为和业态[4]。大数据时代学术传播主要是通过网络平台,以学术成果信息为内核挖掘其背景信息及其相关信息,打破传统学术期刊与数据库的数据孤岛局面,增强数字信息使用分析与二次开发能力,利用数据挖掘技术充分释放文献与数据的功能,把有意义的每一条数据及其数据关系都转换成一个知识群或信息链,增强优质内容的增殖与衍生能力,创造更高的附加值[1]。

1.3　数据挖掘的意义　当今时代,学术期刊办刊人必须对数据具有基本的辨识能力,对数据进行诠释、组织与整理,进而通过数据资源组合产生的能力。有学者提出,纸质书报刊的数据化价值的产生,是大数据技术应用于新闻出版业的初衷和归宿,也是新闻出版业由数字出版向数据出版转型和过渡的关键和标志[7]。

在网络环境中,信息加工的需求为编辑提供更广阔的发展空间,通过对本研究领域大数据相关研究信息的把握,甚至介入信息数据的大数据研究中,掌握第一手资料,及时进行选题策划,可占领学术研究的制高点。早在本世纪初,时任《临床误诊误治》杂志主编的陈晓红主任医师即敏锐感到文献数据中可能隐藏的客观规律,遂带领医学专家、计算机专家和编辑专家团队,在已奠定的误诊研究理论框架基础上,研发出我国第一个临床医学专题文献数据库"误诊疾病数据库检索与管理系统"。

2　误诊大数据对文献数据价值的挖掘

2.1　误诊疾病数据库研究背景　误诊是医生经过临床诊断后,其结论与疾病的本质不符合或不完全符合的现象,是医生在认识疾病过程中期望认识其本质而实际与本质偏离,或仅接近本质的现象。误诊是临床医学永恒的话题,人类对任何疾病的认知,都是从误诊开始的。尽管临床诊断技术不断发展,但疾病总体误诊率并未随之下降。误诊率不变的现象提示,传统诊断学的研究方法可能无法解决避免误诊的问题;于是,我们提出逆向思维的方法,从收集误诊病例做研究,探讨误诊发生的规律,寻找防范的措施,研究误诊的目的在于最大限度地减少误诊[8]。

2.2　误诊疾病数据库研发过程　误诊疾病数据库研究始自 1988 年,首先通过专著《误诊学》和数十篇系列误诊理论研究文献,创立了误诊研究应用理论,明确了误诊定义,规范了误诊后果等级、误诊原因等,制定了"标准误诊文献"的 5 项内容,即文献中有明确的疾病诊断标准、有误诊率、误诊原因、误诊后果、误诊范围等。2006 年起,开始收集误诊病例大数据做预测,对 2003—2014 年发表在全部中文医学期刊中的 500 万篇文献进行关键词和题名双重检索,命中误诊文献6.6 万篇,经专业人员阅读分析全文,共遴选 2.7 万篇标准误诊文献,对符合标准的误诊文献由数据小组人员填制"标准误诊文献卡"后,将采集的数据录入数据库。每篇文献平均采集 120 项数据,包括疾病信息、患者信息、作者信息、期刊信息等[6]。

本项目的理论核心为误诊理论框架体系,数据层为多种原始数据加工清洗后的误诊文献数据、临床病案数据和交互性平台数据,目前已经完成理论研究阶段、数据存储分析和产品设计阶段,将散落在文献数据库中的单篇文献中大量非结构数据,通过标化转化为结构式数据加以统计分析。该数据库的研发、数据库构建、数据收集和存储检索平台建设,均由山东康网网络科技有限公司承担,已于 2015 年取得中华人民共和国计算机软件版权登记证书(第 0293544 号)。

2.3　误诊疾病数据库研制的意义　大数据预测功能的发挥需要通过统计分析来实现,但统计分析并不等同于大数据本身。通过离线计算、分布式计算等计算组件

所统计分析出来的二次数据,才是大数据的精华和核心,是在原有海量数据基础上的价值提升和再发现[6]。出版物文献的数据化价值是指在数字化、碎片化的出版物的基础上,对相关资源进行多维度、立体化知识标引,在于海量数据背后的隐藏价值和潜在价值[7]。

在大数据时代,以病例研究为基础的研究将会受到关注,并提升到一个新的高度,很多分析都将基于患者诊疗全程的数据进行统计分析[2]。在未来高度发达的信息化时代,必须培育临床医生运用大数据决策指导临床工作的习惯。误诊大数据正是将浩如烟海的零散的误诊文献进行深入分析和挖掘,收集设定时间段内的误诊文献全数据,以决策树的形式,对收集到的规范、清洗和标化数据进行纵横交错的统计分析,得出海量数据分析结果,对疾病误诊进行有意义的预测。

3 误诊数据库论文的选题策划

数据库从2006年开始正式录入文献数据,到2014年,已经收录了2003—2014年发表在全部中文医学期刊中的标准误诊文献2.7万篇,累计病例近100万,纠正误诊后的确诊疾病2 400种,越来越多的疾病呈现了规律性结果。为了及时将这些数据结构呈现出来,本刊编辑部着手进行误诊数据论文的选题策划,策划过程如下。

3.1 筛选病种 误诊疾病数据库纳入2 400余种确诊疾病,在众多的疾病中必然有部分最能呈现误诊发生规律的病种,将这些疾病及时加以整理报道,将对临床医师减少误诊和提高诊断水平有重要意义。我们按照误诊病例较大、误诊率较高的原则遴选200种疾病纳入组稿计划,同时遴选误诊病例样本虽较少,但是当下迫切需要临床医师提高认识的新病种,以期前瞻性地干预从而减少该疾病的误诊。

3.2 统计数据 数据库的在线检索系统可有几十项统计结果,包括疾病信息、作者信息、期刊信息等。那么,哪些统计最能反映疾病整体误诊概况呢?我们选择标准误诊文献的5项证据原则,即单病种误诊选出概况、误诊率、误诊范围、确诊手段、误诊后果和误诊原因,认为这些是对于临床医师最有参考价值的精华数据。

3.3 确定体例 误诊数据库论文,不同于寻常的临床研究论文,亦不同于寻常的文献综述,统计结果是来自海量文献的非结构数据生成的结构性数据结构。那么在文稿体例方面,无论是编辑,还是专家,都是首次面对;故必须事先确定文章体例和范式,才能便于作者有的放矢地把握写作思路,也为后期的编校工作打下基础。

3.4 遴选作者 组稿专家的遴选也是一项颇为艰巨的工作。组稿专家既要在本专业领域具有一定影响力且年富力强,还需要对误诊研究特别支持和理解,且其学科团队必须有一定科研实力。笔者在本刊编委队伍的基础上加以拓展。笔者走访了10余座城市20多所医院,同时在各专科学术会议和专科讨论微信群中发现作者,向每一位专家介绍误诊疾病数据库项目,文稿的写作要求等等。最终,我们在全国范围内确定26个专科的知名学者担任组稿人,共有268位临床医师参与了误诊数据库单病种误诊研究的文稿撰写工作,目前已完成202个单病种误诊研究文稿的组稿工作,为误诊大数据选题储备了充足的稿源。

3.5 选题编发 当第1批约稿如此而至,我们从中看到了许多有意义的内容,感到欣喜;故在组织全体编辑人员对文稿进行规范化编校处理的同时,根据期刊重点专栏选题计划酌情安排刊期。2014年第4期至2016年第7期已刊登或纳入编校计划共12篇单病种误诊数据研究[9-20],这标志着国内第1批应用临床大数据研究产出的论著,首次全视野地展示了我国10年单病种误诊概况,从量与量的变化中挖掘其内在的规律。

例如组织国内首批急诊胸痛中心北京大学人民医院急诊科和国家卫计委北京医院急诊科来完成急性心肌梗死和主动脉夹层,分别对1.3万余例急性心肌梗死和6 000例主动脉夹层误诊病例的误诊率、不同级别医院误诊率、误诊范围、误诊后果、误诊原因等进行深入分析。这2类疾病是急性致命性胸痛,呈现了疾病首诊主要症状,最重要的鉴别诊断疾病谱等,呈现了大数据分析的趋势特征。

2016年第1期"急性杀鼠剂中毒误诊专题",邀请中国毒理学会中毒救治专业委员会主任委员邱泽武教授领导的国内最大的中毒急救中心解放军307医院全军中毒救治中心的精锐团队来完成,这是国内首批完整体现杀鼠剂中毒误诊概况和流行病学变化的研究论著。

2016年第2期刊发的应激性心肌病误诊数据分析,我们邀约国内知名的急诊医学专家孟庆义教授组稿,他是最早在文献和自媒体中呼吁临床医师亟待加强对应激性心肌病这一新病种认识的临床专家,在该文中,通过对全误诊文献数据的分析,帮助急诊和心内科医师从急性胸痛患者中及时甄别该病种,拓宽诊断思维,从而发挥前瞻性减少误诊的作用,而这正是我们从事误诊研究的终极目的。

4 误诊大数据选题策划的启示

4.1 临床医学数据研究型论文的尝试 随着信息时代的到来,数据挖掘被越来越多地应用于临床实践。数据挖掘可以对疾病的发生发展提供预测和预警,在公共卫生管理中发挥着重要作用。同时,通过对临床

医学专题数据库信息的挖掘,增加知识更新速度,帮助知识传播,为临床医师决策提供有力的辅助工具。误诊疾病数据库利用计算机技术,对所有数据进行系统归类分析后,在数与数量的关系中发现规律性数据,为临床研究提供有意义的证据支持。误诊数据库产出的海量数据,经过分类统计并深入分析,已经具备了学者们提出的数据论文中数据库论文的主要特质[6],且其数据特质、关联表现形式、数据仓储平台的链接等数据表现形式,技术上已经成熟,这是对我国目前的临床医学数据型论文体例的一种尝试。

4.2　医学期刊大数据出版选题的尝试　传统出版只是转化为数据化出版的数据资源而已,其更大的价值在于,将传统的出版资源进行数字化处理,以网络为平台,实现数据的量化,为资源共享、价值挖掘提供以服务为导向的数据化平台[21]。大数据时代对医学期刊提供了巨大的空间;但由检索和复习相关文献可知,生物医学期刊目前以大数据为主题的报道更多地是人类基因库、公共卫生领域、慢性病管理、传染病预防领域的趋势、发展等述评性文献,少部分报道中涉及临床医学数据库的相关内容[22]。这些临床数据库多为某个学术组织或大医生牵头组织的多中心临床研究所搭建,具有数据来自临床病例的优势;但受体量和收集时间的限制,是否符合大数据的"全数据"特征有待商榷。

5　结束语

出版科学类期刊近年对期刊数据化转型的报道,大多集中在业务流程优化、产品业态和服务模式的设想与展望,对于如何从广袤的大数据研究成果中挖掘选题,尚无更多可借鉴的经验。鉴于此,笔者通过误诊数据库研究选题的策划,无论是医学期刊数据化转型的实践,还是临床医学数据型论文的选题策划,都是较为初步的尝试。谨以粗浅体会抛砖引玉,期待走在数据转型前沿的期刊同道分享更多宝贵经验。

6　参考文献

[1]　夏登武. 大数据时代学术期刊的内容优化与价值重构[J]. 中国科技期刊研究,2016,27(3):264

[2]　吴锋. 大数据时代科技期刊的出版革命及面临的挑战[J]. 出版发行研究,2013(8):66

[3]　周小华. 大数据时代中国学术期刊的转型与发展机遇[J]. 科技与出版,2014(4):102

[4]　赵文义. 学术期刊大数据出版研究[J]. 出版发行研究,2016(3):50

[5]　刘凤红,崔金钟,韩芳桥,等. 数据论文:大数据时代新兴学术论文出版类型探讨[J]. 中国科技期刊研究,2014,25(12):1452

[6]　屈宝强,王凯. 数据论文的出现与发展[J]. 图书与情报,2015(5):1

[7]　张新新. 新闻出版业大数据应用的思索与展望[J]. 科技与出版,2016(3):5

[8]　陈晓红. 中国误诊大数据[M]. 南京:东南大学出版社,2016:3

[9]　赵春菱,杨振,姚丽霞,等. 误诊疾病数据库单病种误诊文献研究:恙虫病[J]. 临床误诊误治,2015,28(4):1

[10]　温伟,张新超. 误诊疾病数据库单病种误诊文献研究:主动脉夹层[J]. 临床误诊误治,2015,28(5):1

[11]　国献素,刘芳. 误诊疾病数据库 2005—2012 年单病种误诊文献研究:小儿气管异物[J]. 临床误诊误治,2015,28(7):1

[12]　周中和,陈会生. 误诊疾病数据库 2005—2012 年单病种误诊文献研究:蛛网膜下腔出血[J]. 临床误诊误治,2015,28(10):1

[13]　刘兆喆,谢晓冬. 误诊疾病数据库 2005—2012 年单病种误诊文献研究:乳腺癌[J]. 临床误诊误治,2015,28(7):1

[14]　董建光,邱泽武. 误诊疾病数据库 2004—2013 年单病种误诊文献研究:致惊厥杀鼠剂中毒[J]. 临床误诊误治,2016,29(1):1

[15]　彭晓波,邱泽武. 误诊疾病数据库 2004—2013 年单病种误诊文献研究:抗凝血杀鼠剂中毒[J]. 临床误诊误治,2016,29(1):5

[16]　何春来,孟庆义. 误诊疾病数据库 2004—2013 年单病种误诊文献研究:应激性心肌病[J]. 临床误诊误治,2016,29(2):1

[17]　李凤鹏,陈会生. 误诊疾病数据库 2004—2013 年单病种误诊文献研究:颅内静脉窦血栓形成[J]. 临床误诊误治,2016,29(4):1

[18]　张秋河,刘芳. 误诊疾病数据库 2004—2013 年单病种误诊文献研究:特发性肺含铁血黄素沉着症[J]. 临床误诊误治,2016,29(5):15

[19]　韩涛,刘兆喆,鞠伶伟,等. 误诊疾病数据库 2004—2013 年单病种误诊文献研究:肾癌[J]. 临床误诊误治,2016,29(6):45

[20]　刘丽萍,余剑波. 误诊疾病数据库 2004—2013 年单病种误诊文献研究:急性心肌梗死[J]. 临床误诊误治,2016,29(7):25

[21]　李德团,雷晓艳. 大数据出版:内涵及其实践应用[J]. 编辑之友,2016(4):23

[22]　秦文哲,陈进,董力. 大数据背景下医学数据挖掘的研究进展及应用[J]. 中国胸心血管外科临床杂志, 2016,23(1):55

(2016-04-17 收稿;2016-06-23 修回)