中国图书馆学报(双月刊)2001 年第 2 期
ZHONGGUO TUSHUGUANXUEBAO

●苏新宁

引文索引数据质量控制研究

摘　要　分析了编制引文索引过程中易出现的数据错误,提出了利用计算机纠正这些错误的思路。这些思路在编制《中文社会科学引文索引》时取得了较好效果。参考文献 4。

关键词　引文索引　数据质量控制　规范文档
分类号　G353.21

ABSTRACT　In this paper, the author analyzes some errors frequently appearing in the compilation of citation indexes, proposes some ideas to correct them with computers, and applies these ideas in *Chinese Social Sciences Citation Index*. 4 refs.

KEY WORDS　Citation index.　Data quality control.　Authority file.
CLASS NUMBER　G353.21

引文索引数据的质量直接关系到引文索引系统的整体效率和它的权威性,它将对引文的统计分析产生极大影响。我们研制《中文社会科学引文索引》时,针对易出现的数据错误,利用软件进行了多方预防和有效纠正,大大减少了人工校对的工作量,确保了引文索引的数据质量。

1　可能产生的数据错误及影响

引文索引的数据主要分两大类:来源文献数据和被引文献数据[1]。

来源文献的数据项主要包括:论文标题、作者、机构名称、部门名称、地区、期刊名称、论文发表年代、卷期页码、引文数量以及由人工标引的文章类别、作者类别(个人或团体)、机构类别、分类号、标引词等[2]。

被引文献的数据项主要包括:被引文作者、篇名、被引文献形式(论文、图书、报告、标准、专利等)、引用类别(参考文献、页注、文内注等)、被引期刊名称、被引文年代、引用期刊代码、非期刊论文被引出处等[3]。

1.1　引文索引数据易出现的错误

引文索引的数据,每一项都有可能在输入时产生错误。当然,这可以通过双轨输入比较出差异(错误),从而纠正它们。但由于引文索引中数据项多,符号繁杂,出现的差异非常之多,许多差异对于数据的质量是无关紧要的(如符号、数字和空格的半角与全角,西文的大小写等)。这些差异都需要人工去鉴别和修改,这样既花费了双倍的录入量,也增加了人工鉴别与修改的工作量。对于这样复杂的数据录入,我们仍采用单轨输入的方式。在实践中我们发现,除了录入员可能产生错误外,作者本人、杂志的印刷等也都可能出现错误,许多错误如果缺乏比较很难发现。

来源文献的数据错误主要有:

(1)文字上的错误。有可能是来自刊物本身的印刷错误,也可能是录入时产生的错误。

(2)机构名称不规范。如:"北京大学"写成"北大",将"安徽省社会科学院"著录成"安徽社科院"等。

(3)机构与机构类别不匹配。如:把"北京大学"著录成"科研院所类"机构等。

(4)把部门名称当作机构名称著录。如:把"南京大学商学院"当作机构来著录,而它的机构应是"南京大学","商学院"则是部门名称。

被引文献的数据错误更多,绝大部分错误出自作者或编辑部。主要表现如下:

(1)被引期刊名称不一致。期刊名的简写和缩写很多,有的作者完全是根据自己的想象给出刊名。如:《复旦学报》被写成了《复旦大学学报》,《北京大学学报》被改成《北大学报》等。

(2)被引论文标题不一致。有的标题被"裁剪",有的标题被丢字,造成同一篇文章的标题被写成多种多样。例如,费孝通先生 1980 年在《中国社会科学》

第 1 期上发表的论文《关于我国的民族识别问题》,在 1998 年被引用 4 次,但该文的题名就有 4 种不同写法。

(3)被引论文作者姓名被错误标引。如:作者"黄宗忠"变成了"黄忠宗","马费成"被输成了"马费城"等。

1.2 数据错误带来的影响

以上错误除了影响引文索引的查全率和查准率,更重要的是将使人们对引文索引的各类分析统计结果产生偏差。例如,机构名称的错误将影响到机构发文量的统计;机构类别的错误,将会把一个机构分别在不同类别的机构中统计;引文期刊名称的错误,将会导致期刊被引次数统计出现误差,从而影响各学科期刊的排队、期刊影响因子的计算以及核心期刊的确定;论文作者姓名的错误将使对作者的发文和被引情况的统计产生误差,从而影响到对各学科核心学者群的确定;当标题出现错误时,就可能使人们对重要论文的评价产生误差。我们在进行论文被引统计时,常常发现实际是同一篇文献,却由于篇名不同被统计在两处或更多处。这就影响到对各学科核心论文的分析和评价。

2 纠正错误的思路

上述错误直接影响到引文索引的查询效率和统计结果,在引文索引正式发布前,必须纠正这类错误。我们首先考虑的是人工仔细校对,但后来发现许多错误单凭人工,很难发现。例如,刊物本身的印刷错误,作者自身的错误和简写等。我们也考虑过预先建立规范字典(机构字典、期刊字典等),但凭个人想象建立的规范字典很难覆盖各类错误(输入错误、刊物印刷错误、作者错误)。能否用计算机实现自动校对?计算机校对能否提高校对率?计算机校对的可靠性如何?这些都是我们一直在思考的问题。经过对数据的研究分析,我们认为,计算机完全能够替代人进行数据校对。通过实践检验,计算机自动校对引文数据确实达到了令人满意的效果。

我们所采取的校对思路如下:

首先,规范并统一来源文献作者机构的名称。确保对机构的统计分析数据正确,保证对机构查询具有高查准率和高查全率。

第二,机构名称与机构类别的对应。机构发文量的排队是根据机构类别进行的,机构名称和机构类别的统一,是确保机构正确进行发文统计的又一保证。

第三,规范被引期刊刊名。利用引文索引确定核心期刊的一个主要途径是期刊被引数量,并由此得到期刊的影响因子。被引期刊刊名的不规范直接影响到对期刊影响因子的计算。

第四,被引文献作者的姓名纠正。发文作者的粗心或录入员的错误,可能导致被引文作者姓名被错误输入,这可能影响到对作者被引量的统计结果,使各学科的核心学者的确定出现误差。

第五,被引文献篇名的更正。在对被引文献篇名的分析中我们发现,原稿中错误的比例非常大。如:少或多一个"的",把篇名缩减了;作者凭记忆写出引文篇名等。

3 纠正错误的方法

所有的校对实际上都可以直接用人工进行,但考虑到效率以及正确率,人工校对工作量大,有的原稿错误人工根本无法校出,因此,利用计算机进行高效准确的校对则是我们的最终目标。我们努力将以上校对思路实现计算机化,实践证明,它是行之有效的。我们具体采取的方法与算法如下。

3.1 机构的自动校对

发文机构的错误主要表现在:用简称代替全称;机构名称不完全;将部门名称加入机构;录入员的输入错误和机构类别的标错等。为纠正这些错误,我们主要采取 3 个步骤。

3.1.1 创建规范字典

建立机构规范字典数据库结构。字典内容包括:规范名称字段、不规范名称字段、机构类别字段。利用软件将引文索引中现有机构名称取出后归并,然后写入机构规范字典的规范名称和不规范名称两个子段。

3.1.2 完善规范字典

用人工对初建的机构规范字典进行校对。主要任务:更正每条记录中的规范名称字段,同时给出机构类别代码。例如,"北京师范大学"被著录成"北师大",而机构规范字典中该记录的两个机构名称的初始值均为"北师大",这时可通过键盘将该记录的"规范名称字段"内容改为"北京师范大学"。采取同样的方式再将机构类别改正。

3.1.3 更正作者机构

利用软件以批处理方式实现对来源文献的作者

机构自动更正。采取方法:顺序从规范字典中取出记录,用不规范名称字段数据对来源文献数据库的机构字段进行搜索,用字典中规范名称字段数据去覆盖所有命中记录的机构字段,用字典中机构类别字段数据去覆盖所有命中记录的机构类别字段内容。该软件运行结束后,基本可以保证来源文献的作者机构信息正确和统一。

3.2　被引刊名的校对

被引文献的期刊刊名校对,其方式和手段与作者的机构校对基本相同。所建立的被引期刊规范字典包含 3 个子段:期刊规范名称、期刊非规范名称、被引文献形式代码。操作过程与机构校对基本相同,以下几个方面与机构校验稍有区别。

(1)期刊刊名归并写入期刊刊名规范字典时,将被引文献形式均默认为是期刊论文类型。

(2)人工校对完善期刊规范时,对非期刊刊名的记录(书名、会议名、出版社等),更正其被引文献的形式代码。

(3)以批处理方式规范被引论文期刊名时,同时将被引文献形式更正。若字典当前记录的被引文献形式为非期刊论文,需将命中文献的类型更正,同时将刊名字段中的数据移至相应字段[4]。

(4)将期刊刊名规范字典中的非刊名记录剔除,以保证规范字典的"纯洁"。

3.3　被引作者和篇名的校对

被引文献的期刊刊名校对后,可进一步校对被引文献的篇名和作者。校对采取交替比较的方式进行。

例如,作者、刊名、卷期相同,而篇名不相同者,将不同情况列出由人工确定正确者。对同一作者、刊名和卷期,有多条篇名相同,只有一两条篇名不同者,则自动更正。也可比较篇名、刊名和卷期来更正作者名。

篇名较作者长,出现错误的可能性也大,故通常采取先校验篇名,更正篇名后再校验作者。

篇名、作者的一轮校对完毕后,再进行一次模糊匹配校对,把非常相似的被引记录列在一起,进行最后一次人工核对,把数据错误控制在最小范围。

对篇名和作者的校对过程中产生的校对纠正文件只对当年数据有效,并不作为规范字典保存。

3.4　规范字典的利用与维护

初次生成的"机构字典"和"期刊字典",将在以后引文索引的校对中发挥作用,当输入了部分数据后,用规范字典去校验并更正。规范字典中尚不存在的机构名或期刊名,将它们列出并及时补充加入规范字典。

4　结论

以上对引文索引数据质量控制进行了探讨。这些设想也都在我们所进行的《中文社会科学引文索引》研制中得到了实现,并获得了非常好的校验效果和较高效率。这为我们进行机构的发文统计和排队、期刊的被引统计和影响因子的计算、学者被引情况的统计和分析、重要论文的分析等提供了较为可靠的数据保证,也为今后引文索引的数据校验提供了切合实际的规范字典。

参考文献

1～4　苏新宁. 中国社会科学引文索引设计. 情报学报, 2000,19(4)

注　本文为教育部"九五"重大课题成果。项目代号:99JBZD870001。

苏新宁　教授,博士生导师,主要从事情报检索算法、信息处理技术以及网络信息资源的研究与开发等方面的研究。通讯地址:南京市南京大学信息管理系。邮编 210093。

(来稿时间:2000-08-31)