Digital Library Forum 数字图书馆论坛
http: // www.dlf.net.cn

NSTL外文期刊引文数据自动化拆分的研究与实践

鲜国建 赵瑞雪 金晨／ 中国农业科学院农业信息研究所 北京市 100081

摘要：文章简要分析了NSTL国际科学引文数据库的建设现状，讨论了期刊类型引文数据自动化拆分的必要性和可行性，深入研究了期刊类型引文的著录规律，提出采用分类的思想将引文数据划分为不同类型再分别加以拆分，设计出自动化拆分的具体流程和技术框架，实现了自动化拆分工具，并在农学领域进行了自动化批量拆分的应用实践，增强了大规模数据的自动化处理能力，提高了数据的整体质量及时效性。

关键词：NSTL，国际科学引文数据库，引文数据，自动化拆分
DOI：10.3772/j.issn.1673—2286.2010.10.019

引言

引文即文献末尾所附的参考文献，它是文献的必要组成部分。引文客观地描绘出科研工作中文献的利用关系，揭示科学研究及其成果之间的传播、借鉴、继承和发展的脉络[1]。通过对引用的参考文献检索，将全文文献与引用的参考文献链接起来，可实现基于参考文献的源文献查找，实现文献所述知识点的追根溯源，是拓展信息资源、提高获取知识效率的有效方法[2]。此外，引文还是评价科研人员和科研机构学术地位的重要指标。因此，NSTL高度重视引文数据库的建设，自2006年开始投入建设国际科学引文数据库（Database of International Science Citation，简称DISC）。DISC是集文献发现、引文链接、原文传递为一体的服务系统，为我国科研人员提供世界科学研究的脉络，为方便他们了解世界先进国家研究的动态与研究方法提供了重要途径[3]。本文从微观的角度探讨了在建设DISC的过程中，如何改变以手工为主的加工方式，提高引文数据加工的自动化水平，缩短加工周期，提升引文数据的整体质量。

1 引文数据库建设现状

到目前为止，DISC共收录全球出版的3000余种核心期刊，覆盖了理、工、农、医、标准和计量各领域。从2006年开始建设至今，DISC已累积加工建设引文数据3000余万条，且这个数字在不断快速地增长。以中国农业科学院农业信息研究所（以下简称农科）承担的农学领域为例，2009年共完成了300余万条的加工任务。在NSTL引文数据库著录规则中明确要求，揭示的参考文献种类要全面，包括期刊、图书、会议、标准、学位论文、网络资源、专利文献、科技报告以及其他类型；著录的字段要准确翔实，包括引文的类型、作者、题名、出处、年、卷、期、页码、出版地、出版公司、主编和网络信息等多个著录项；加工语种除了英、法、德等主要西语语言外，还要涵盖多种小语种文献。可以说，引文数据加工的数据规模巨大，文献种类繁多且语言类型复杂[4]。因此，NSTL各成员单位花费了大量的人力、物力和时间，以此来保证顺利完成各自的引文数据加工任务。

目前大多数加工单位在数据加工过程中，仍以人工方式为主，计算机为辅，这种加工方式不但效率较低，且质量难以保证，人员培训成本也较高。当前各单位的加工任务已基本处于饱和状态，一旦需要增加加工品种，目前的加工方式则将难以应对。因此，各加工单位有必要进一步优化改善现有的加工流程，改变以人工为主的局面，充分发挥计算机的性能和优势，确保能按时、按质、按量完成加工任务，且在承担更多的加工

任务方面留有余地。

2 引文自动化拆分的研究

2.1 必要性和可行性

每年千万余条的数据加工任务，使得NSTL各成员单位都安排多位数据加工人员参与其中。由于大部分工作都是通过人工来完成，因而数据加工周期长、效率低，成本较高，数据质量也参差不齐。因此，为了缩短数据加工周期，增强数据时效性，提高数据整体质量，有必要积极挖掘海量引文数据中潜在的规律，为实现计算机自动化批量加工数据提供依据。

表1 引文数据中期刊、图书等类型比例分布

任务批次	引文加工量	期刊类型/百分比	图书类型/百分比	其他类型/百分比
批次1	559790	457084/81.6%	36901/6.6%	65805/11.8%
批次2	516007	414550/80.3%	37692/7.3%	63765/11.4%
批次3	378426	311035/82.2%	26850/7.1%	40541/10.7%
批次4	455422	380390/83.5%	27692/6.1%	47340/10.4%
批次5	355137	294629/83%	25364/7.1%	35144/9.9%

尽管从表面上看海量的引文数据显得杂乱无章，但只要认真深入分析，仍能发现其实还是有章可循的。首先，从对农科已加工的引文数据按期刊、图书等参考文献类型进行统计分析得知（如表1所示），各种类型所占的比例大致为：期刊类型约占82%，图书类型约为7%，会议、学位论文和专利等其他类型共占11%左右。每一批次的加工结果基本都符合这样一个比例分布。可见，期刊类型的引文数据所占比例较大。另外，引文字段的著录即引文拆分的工作量约占在整个加工流程的50-60%。由此可见，如果能解决期刊类型引文数据的自动化批量拆分工作，将大大提高引文加工的整体效率。通过对期刊类型引文数据作进一步的分析和研究发现，这些数据中同样隐藏着一些重要规律，这为实现计算机进行自动化批量拆分提供了可能。

2.2 期刊类型引文著录规律分析

长期以来，分类法都是一种认识事物、区别事物的重要方法。同样地，通过对大量期刊类型引文数据的深入分析发现，这些数据在著录规范方面也蕴藏着很多有价值的规律，以此为依据可将这些数据划分为若干种类型。下面以农科加工的期刊类型的引文数据为例，详细介绍几种最常见的类型。以下是几条样例数据：

（1）Zubair, A.R., A.S. Munir & S. Ahmad 2007. Efficacy of different insecticides against sugarcane termite (Microtermes spp.). Journal of Agricultural Research 45(3):215-219.

（2）Parker KL, Robbins CT, Hanley TA (1984) Energy expenditure for locomotion by mule deer and elk. Journal of Wildlife Management, 48, 474-488.

（3）Hedley et al., 1999 S.L. Hedley, S.T. Buckland and D.L. Borchers, Spatial modelling from line transect data, Journal of Cetacean Research and Management 1 (3) (1999), pp. 255-264.

（4）M. Ishihara, M. Matsunaga, N. Hayashi, V. Tisler, Utilization of D-xylose as carbon source for production of bacterial cellulose, Enzyme Microb. Technol. 31 (2002) 986-991.

（5）Dodd IC, Stikic R, Davies WJ. Chemical regulation of gas exchange and growth of plants in drying soil in the field. Journal of Experimental Botany (1996) 47:1475-1490.

从上面的实例数据中不难看出其中的著录规范，第一条数据的规范为引文作者之后紧跟引文年，引文年与出处之间为题名（可能没有），出处之后是引文卷和期（若有期的话，则是放在一对圆括号内），最后是引文页码，引文页码与卷期之前以冒号分隔。即类型一的著录规范可表示为：“作者 + 年份 + 【题名】+ 期刊名称 + 卷 + 【（期）】+ ; + 页码”，其中的期刊名称即为出处，题名和期两边的方括号表示它们可能有或没有。

依此类推，可得出其他几种类型的著录规范为：

类型二：“作者 + (年份) + 【题名】+ 期刊名称 + 卷 + 【（期）】+, +页码”；

类型三：“作者摘要 + 年份 + 作者详情 + 【题名】+ 期刊名称 + 卷+ 【（期）】(年份) + 页码”；

类型四：“作者 + 【题名】+ 期刊名称 + 卷 + 【（期）】+ (年份) +页码”；

类型五：“作者 + 【题名】+ 期刊名称 + (年份) +

卷+【（期）】+页码”。

上述几种类型引文数据的区别主要在于，类型一中年份两边没圆括号，类型二及其他类型中都有；类型三有两个显著特征，一是年份出现两次，且第二次有圆括号，二是先列出作者摘要，然后在第一个年份后列出作者详情；类型四的特征主要是年份在卷期与页码之间，而类型五的年份则处于期刊名称和卷期之间。通过对大量数据的对比分析发现，绝大部分期刊类型引文数据都遵循上述著录规范。通过著录项的著录位置和著录次数以及圆括号、冒号等这些特征信息，即可实现将95%以上的期刊引文数据划分成不同的类型。针对每种类型设计相应的拆分算法，将能实现这些类型数据的自动化批量拆分。通过这种分类的方式来拆分数据，既能减少各类数据之间的干扰，又能提高拆分的效率和准确性。

2.3 拆分流程和人员配备

此处的拆分流程主要是以包括期刊、图书、学位论文等各种类型的原始引文数据装载到数据库之后为起点。要实现对期刊类型引文数据的批量拆分，首先是需要将非期刊类型的数据筛选出去，而区别期刊和图书等非期刊的主要依据就是卷期信息和一系列的标志词，如图书的Publisher和Publishing house、会议的Proceeding和symposium、学位论文的Ph. D. Thesis等。因此，拆分的首要工作就是将引文数据中的年、卷、期等信息拆分出来。通过对比分析发现，年份、卷、期和页码的著录也有一定的规律，通过一定的方式也能完成批量拆分，实现起来相对简单。与此相比，作者、题名和期刊名称（即出处）的拆分要复杂得多，因此需要设计更为准确、细致的拆分算法来完成。具体的拆分流程如图1所示。

通过拆分出来的年、卷、期、页码和一些特征词，可将所有的引文数据拆分为期刊、图书、学位论文等大类。期刊类型数据则可以进一步细分为多种子类型，针对每种子类型再进行作者、题名和出处的拆分，其中作者部分还需要提取第一、第二和第三作者。图书等非期刊类型的数据以及少量未拆分的期刊类型的数据规律性不强，用计算机自动拆分的难度较大，主要还是以人工拆分为主。人工拆分和计算机批量拆分都难免有错误，还需要对所有拆分后的数据进行审核和修正。

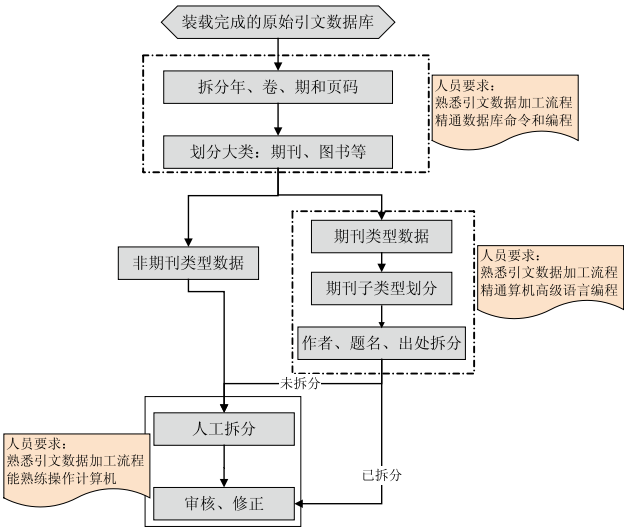

图1 期刊类型引文数据拆分流程图

人员配备方面，由于在整个拆分过程中，期刊、图书等大类的划分、期刊子类型的划分及各子类型的作者、题名和出处的拆分技术性较强，对加工人员的要求较高，不但要熟悉引文加工的整个流程，还要精通数据库命令和丰富的计算机编程经验，这样的技术人员需要1-2名。而对参与人工拆分、审核和修正的加工人员没有太高要求，只要能熟悉加工流程、能熟练使用计算机、有认真负责的态度就能胜任，配备这类加工人员的数量可视承担的任务量而定。

2.4 自动化拆分的技术框架

要实现期刊类型引文数据的自动化批量拆分，可采用如下图2所示的技术框架。

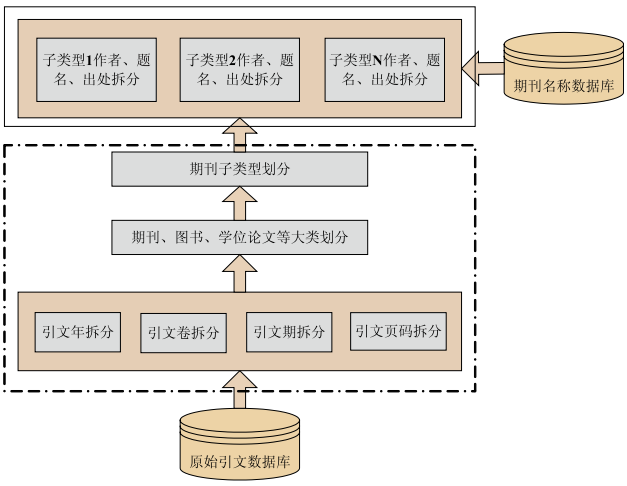

图2 自动化拆分的技术框架图

其中，虚线框部分主要负责年、卷、期和页码的拆分，这将为拆分期刊、图书等大类和进一步拆分期刊子类型提供关键特征信息。这部分操作相对简单，可直接在操作命令比较丰富的数据库（如FoxPro）中完成，为了能重复使用也可编写一些拆分程序。

实线框部分则主要完成期刊子类型中作者、题名、出处字段的拆分工作。这部分拆分需要有一个关键的期刊名称数据库（简称期刊库）做支撑。所谓的期刊库就是指由期刊类型引文数据中的期刊名称（即出处）组成的数据库。通过大量实验发现，在拆分题名和出处时，简单地以字母大小写或句点等标点符号很难准确定出处的分隔位置，而如果以期刊库中的期刊名称去查找引文库中有此刊名的引文数据并加以拆分，则可大大降低拆错的比率。

在此技术框架中，除了各种拆分算法需要精心设计和不断优化外，期刊库也需要不断地丰富和充实。只有期刊库中收集的刊名越全，能被正确拆分的引文数据才会越多。还需要注意的是，为了尽可能地减少拆分不准确的情况，需要跟年、卷、期等配合使用。此外，在从期刊库抽取刊名时，最好从长度最长的刊名开始。比如刊名为“Science”和“Danish Journal of Plant and Soil Science”，则最好先从第二个刊名开始，不然就可能会截断而导致拆分不准确。

3 引文自动化拆分的实践

农科在深入分析农学领域的引文数据著录规范规律基础上，采用分类的思想，实现了上述的技术框架。其中，框架中的虚线部分主要是在FoxPro数据库中完成。由于FoxPro数据库操作命令丰富，使用简单灵活，加上编写一部分程序，已基本实现年、卷、期和页码的拆分和期刊、图书等大类的划分，以及期刊子类型的进一步划分。而期刊子类型中的作者、题名和出处则是通过在Visual Studio .Net 2005平台下，以期刊库为基础，针对各子类型编写较为复杂的拆分程序来完成拆分工作。图3就是拆分期刊子类型的操作界面。

考虑到引文中的多个作者之间的组合方式也非常复杂，因此在.Net平台中只负责把所有的作者作为一个整体拆分出来。从中提取第一、第二和第三作者的工作则留在FoxPro中进行，仍利用一些关键特征对其加以分类，然后再进行拆分。

图3 拆分期刊子类型的操作界面

此外，我们还对自动化批量拆分的结果进行认真的分析，找出拆分错误的数据并分析了出错的原因及规律，有针对性地改进拆分算法，并编写了相应的质检程序和批量修复程序，如图4所示。通过这些措施，基本能确保将期刊类型引文数据自动化批量拆分的准确率控制在98%左右。同时，也高度重视期刊库的质量和数量，每一批次加工任务完成过程中，都将因期刊库不全而导致未被拆分的引文数据中的刊名提取，审核后加入到期刊库中，使得期刊库处于不断丰富和充实的状态。

图4 批量修复数据

4 结束语

本文通过分析期刊引文自动化拆分的必要性和可行性，深入研究期刊类型引文的著录规律，提出了具体的拆分流程以及相应的技术框架。农科结合承

担的加工任务，在期刊类型引文数据进行自动化批量拆分方面进行了尝试。实践结果表明，这不仅大大提高了数据加工的效率，缩短了加工周期，数据的整体质量也得到了进一步提高。然而，在实践中也还发现存在的一些不足之处，如拆分过程中需要人工干预的环节还比较多，对加工人员的计算机技术能力要求比较高。随着期刊库规模的扩大，仅采用期刊名长度优先的方式会导致拆分效率的降低。在后续的研究和实践，还需不断优化流程，在确保准确率的前提下进一步提高拆分效率，减少人工干预环节，提高批量拆分的自动化水平，并加强质量控制和流程管理。

参考文献

[1] 袁培国,等. 论引文索引数据用作评价工具的科学性和局限性[J]. 学术界,2009(3):47-56.
[2] 方义,黄胜海. 科技文献中引文的功能应用和检索方法[J]. 农业网络信息,2008(10):50-52.
[3] 国际科学引文数据库[DB/OL]. [2010-03-20]. http://citation.nstl.gov.cn/index.jsp.
[4] 任慧玲,等. NSTL国际科学引文数据库医学外文期刊引文数据加工流程和加工技术研究[J]. 医学信息学杂志,2009,30(3):19-21.

作者简介

鲜国建（1982-），男，硕士。研究方向：叙词表、本体、数字资源加工、信息系统开发。通讯地址：北京市中关村南大街12号 100081。E-mail: xgj@mail.caas.net.cn
赵瑞雪（1968-），女，研究员，博士生导师。主要研究方向：信息管理与信息系统的理论、方法和技术，数据库应用技术与方法，信息系统集成，发表学术论文30多篇。

Study and Practice on Automatically Splitting of NSTL's Foreign Journals' Citation Data

Xian Guojian, Zhao Ruixue, Jin Chen / Agricultural Information Institute of CAAS, Beijing, 100081

Abstract: This paper gives brief introduction about the construction of Database of International Science Citation (DISC), and makes some discussions on the necessity and feasibility to split the journals' citation data automatically. We make further study on the marked rules of citation data with journal types, and propose to classify the citation data into different types such as journals, books and so on, and then split them respectively. We design the workflow and technical framework, and also develop a system to classify and split the citation data automatically. The practice on agricultural domain proves that this system enhances the processing capability with large-scale data and also improves the effectiveness and the whole quality of DISC.

Keywords：NSTL, DISC, Citation data, Automatically splitting

（收稿日期：2010-08-30）