

基于 ISI Web of Knowledge 引证检索 服务统计软件设计与实现

马海收^{1,2} 刘媛媛^{1,2} 郑 菲¹ 谢华玲¹

(1. 中国科学院国家科学图书馆 北京 100190;
2. 中国科学院研究生院 北京 100049)

摘 要 结合国内图书馆普遍开展的论文引证检索服务的实际需求,在大量工作实践的基础上,设计并实现了一款基于 ISI Web of Knowledge 平台检索结果引证检索统计报告的软件,能够根据不同的统计指标,对检索结果进行快速统计。实践证明,该软件提高了工作效率的同时,保证了正确率。

关键词 ISI Web of Knowledge 引文统计 软件设计

中图分类号 G350.7

文献标识码 A

文章编号 1002-1965(2012)02-0148-05

Design and Implementation of Statistics Software Based on ISI Web of Knowledge Cited Reference Search Service

MA Haishou^{1,2} LIU Yuanyuan^{1,2} ZHENG Fei¹ XIE Hualing¹

(1. National Science Library, Chinese Academy of Sciences, Beijing 100190;
2. Graduate University of Chinese Academy of Sciences, Beijing 100049)

Abstract To meet the actual needs of cited reference search service provided by domestic libraries, on the basis of plenty of practical work, we design a software based on citation report in the search results of ISI web of knowledge. Being able to make quick statistical analysis of search results by different statistical indicators, this software proves to be both efficient and accurate.

Key words ISI Web of Knowledge quotation statistics software design

0 引 言

论文被引次数很大程度上反映了该论文影响力的大小,而科研工作者的论文的被引次数则说明了其研究成果在其研究领域内的影响力。通过对引文的统计分析可以评价科研工作者及其研究成果,科研团队和科研项目研究成果,也可以评价国家、地区及科研机构的科研水平^[1]。被引频次包括他引和自引两个基本部分,由于自引的特殊性,需要将自引频次中的自引排除,他引能更加准确、客观地反应出论文的影响力和科学家的研究成果。在现实工作中,往往需要使用他引来评价科研工作者及其研究成果。由于目前部分论文的合作者人数众多,如果施引论文中任意一个作者出现在被引论文中则视为自引,人工排除自引的工作量

非常大,且容易出错。

ISI Web of Knowledge 是汤森路透科技信息集团开发的产品,是为科研人员建立的研究平台。平台提供复杂的检索功能,并可以查看每一条检索记录的全文、参考文献、被引频次、相关记录和其他信息^[2]。本文利用该平台的被引参考文献搜索结果,设计并实现了一款可以自动排除自引的软件,软件可以帮助工作人员统计科研人员的他引次数,以一定的格式输出到表格中,方便工作人员进行后期的人工排查错误,大大提高效率的同时,也保证了正确率。

1 项目需求和设计思路

1.1 项目需求 在很多情况下需要对科研人员、项目组的研究成果进行客观的评价,而发表的论文数量

以及论文他引情况是研究成果的客观数字化的证明,例如国家杰出青年基金的申报需要统计科研人员在一定年限内发表论文的引用情况,科研项目结题时统计该项目组发表论文数及其引用情况用以评价该项目的研究成果。通过公开发表的科技论文被国际、国内著名索引数据库收录以及引用检索和统计分析,为个人、团体等的学术水平进行文献计量学评价。根据委托人提供的信息如姓名、单位、期刊名称、会议名称、论文题名、发表时间等,查找论文被选定数据库收录及被引用情况,并依据检索要求和检索结果出具检索证明或统计分析报告。

在生成报告过程中,利用 SCI 数据库查询申请者每篇论文的被引用情况,手工分析引用文献,确定他引的结果。这一过程需要进行大量的手工操作,花费大量的时间和精力。为提高分析的工作效率,本文在大量工作实践的基础上,设计了一款协同制作引证检索统计报告的软件,能够根据不同的统计指标,对检索结果进行快速统计。

1.2 功能模块分析 根据实际需求,功能模块分为三个部分,分别为合作者分析、单一作者分析和团体作者分析。三部分功能的主要区别在于排除自引的方法不同。a. 合作者分析:排除自引的方法是排除被引论文的全部作者,即第一作者及其合作者其中任意一人出现在施引论文的作者中即视为自引。b. 单一作者分析:只排除固定一位作者即可,即只有该作者出现在施引论文的作者中才视为自引。c. 团体作者分析:排除固定某几位作者,即人为规定的固定几位作者任意一位出现在施引论文的作者中则视为自引。

以上三个模块的功能基本相同,都是根据一定规则排除自引统计他引。但合作者分析比后面二者功能复杂,因为合作者分析中需要排除的作者是施引论文的全部合作者,要排除自引首先要找到全部合作者,且论文不同其合作者也随之变化。而后两者只是排除固定的作者。

1.3 设计思路

1.3.1 合作者分析设计思路。具体需要统计引用的论文由委托人以简写的形式提供,例如 ZHAN XW 的一篇论文于 2007 年发表在期刊《J AM CHEM SOC》129 卷第 7246 页上,其论文简写信息为“ZHAN XW J AM CHEM SOC 2007 129 7246”,信息构成分别为论文的第一作者名称、发表的期刊简写名称、年卷页信息。首先,从 ISI Web of Knowledge 中下载检索到的数据库收录该作者对应统计年限内的论文信息^[3];其次,分别根据提供的各个论文简写信息,在 ISI Web of Knowledge 的 Web of Science 被引参考文献检索中检索下载施引论文信息^[4]。通过一定的判定方法将论

文简写信息与收录论文信息对应,找到该条论文简写信息的全部作者,将这些作者与施引论文的作者一一进行排查,统计自引和他引数量;最后,将统计结果和自引的作者名输出到表格中,达到统计数据和方便人工检查错误的目的。

合作者分析的设计思路流程图如图 1 所示,具体流程为:a. 检索获取数据源,并将统计的科研人员的英文引用名称作为数据源文本文件的文件名。b. 读入收录论文的信息,包括收录论文的个数、作者、标题、来源、ISSN 等信息。c. 逐项分析提供的论文简写信息,根据信息中的第一作者、期刊名、年卷页数等信息判断与第二步中读入的论文收录信息进行匹配,如果匹配成功则记录下对应收录论文的序号,在排除自引的过程中,将相应收录论文的作者与施引论文的作者一一比较;如果没有匹配的收录论文信息,只能排除被引论文已知的作者,在论文简写信息中可以得到论文的第一作者;此外,统计的该名科研人员肯定是论文的作者之一,为便于判断,将其英文引用名称作为数据源文本文件的名称。因此在这种情况下只排除提供的论文简写信息中第一作者和统计作者的名称即可。找到首个相同作者后,即记录下作者名称并停止比较。d. 判断结束后,输出统计结果到 Excel 表中。

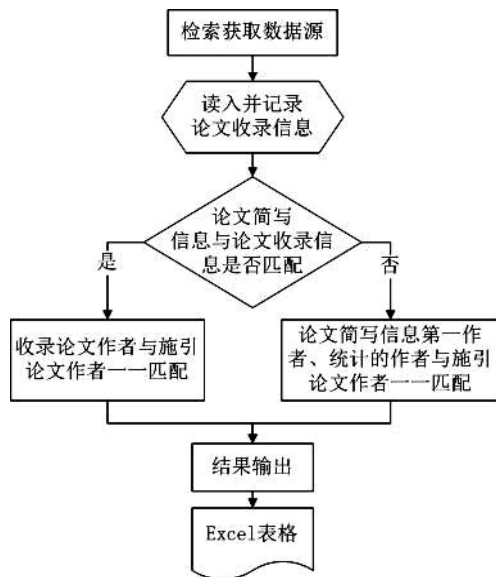


图1 合作者分析总体设计思路流程图

1.3.2 单一作者和团体作者分析设计思路。因为已经确定需要排除的作者名单,不需要找到论文简写信息对应论文的全部作者,因此不需要下载相应数据库收录的论文信息。直接根据提供的各个论文简写信息,在 ISI Web of Knowledge 的 Web of Science 被引参考文献检索中检索下载施引论文信息,存储到文本文件中作为数据源。数据源的命名规则,将需要排除的全部作者的英文引用名称作为数据源的文件名,并将不同作者名用分号隔开。选择数据源后,将施引论

文中的作者与要排除的作者进行比较,统计结果并输出。由于两个模块的思路简单,且可以看作为合作者分析的一部分,设计实现部分只针对合作者分析模块进行详细描述。

2 设计实现

2.1 检索获取数据源 输入数据源包括收录论文信息和施引论文信息两个部分,获取流程图如图2所示。下面对两部分数据获取做详细介绍。

2.1.1 获取收录论文信息。以SCI数据库为例,获取数据来源第一步首先下载SCI收录的包含该作者(可能不是第一作者)的论文详细信息,下载格式以网站标准下载格式为准,其中包含文章的序号(以Record X of X为标志,其中前者为文章序号,后者为收录论文总数)、作者、论文名称、来源、期刊ISSN、DOI等信息。如果针对每一条论文简写信息分别下载其详细信息,会大大增加人工工作量,采用一次下载全部收录论文方式的目的是提高数据采集的效率。但缺点是需要通过相关信息的判断来决定论文简写信息与下载的论文详细信息之间的对应,这个问题的解决会在后面提到。数据示例如下所示:

Record 1 of 22

Author(s): Zhan, XW (Zhan, Xiaowei); Tan, ZA (Tan, Zhan 'ao); Domercq, B (Domercq, Benoit); An, ZS (An, Zesheng); Zhang, X (Zhang, Xuan); Barlow, S (Barlow, Stephen); Li, YF (Li, Yongfang); Zhu, DB (Zhu, Daoben); Kippelen, B (Kippelen, Bernard); Marder, SR (Marder, Seth R.)

Title: A high-mobility electron-transport polymer with broad absorption and its use in field-effect transistors and all-polymer solar cells

Source: JOURNAL OF THE AMERICAN CHEMICAL SOCIETY, 129 (23): 7246-+ JUN 13 2007

ISSN: 0002-7863

ISI Document Delivery No.: 176GH

Redord 2 of 22

.....

另外,可以手动补充收录论文信息。如果检索到的收录论文信息太少,论文简写信息找不到对应的收录论文信息,无法找到全部作者信息,对他引的统计产生不利影响。程序允许用户手动添加论文信息,只需按照以上描述的格式,并在开头记录总共包含的信息条数即可。前面的例子中如果手动添加3条论文信息,在文件第一行添加记录“Record 1 of 25”。

2.1.2 获取施引论文信息。在ISI Web of Knowledge的Web of Science被引参考文献检索中,针对每一篇需要统计引用的论文简写信息,检索下载其施引论文信息,这些信息包括原有的论文简写信息、序号、作者、论文名称和来源信息。下载所有需要统计的施引论文信息,按照网站原有的格式放在收录论文信

息尾部,存放在文本文件中,供程序使用。例如根据作者提供的文章信息如“ZHAN XW J AM CHEM SOC 2007 129 7246”,在被引参考文献检索中,共检索到194篇施引文献。信息的格式如下所示:

ZHAN XW J AM CHEM SOC 2007 129 7246

Record 1 of 194

Author(s): Zhang, C (Zhang, Cheng); Nguyen, TH (Nguyen, Thuong H.); Sun, JY (Sun, Jianyuan); Li, R (Li, Rui); Black, S (Black, Sueley); Bonner, CE (Bonner, Carl E.); Sun, SS (Sun, Sam-Shajing)

Title: Design, Synthesis, Characterization, and Modeling of a Series of S,S-Dioxothienylenevinylene-Based Conjugated Polymers with Evolving Frontier Orbitals

Source: MACROMOLECULES, 42 (3): 663-670 FEB 10 2009

Record 2 of 194

.....

收集所有论文简写信息的施引论文信息,按照以上提到的信息组织方式罗列在第一步收录论文信息的后面。

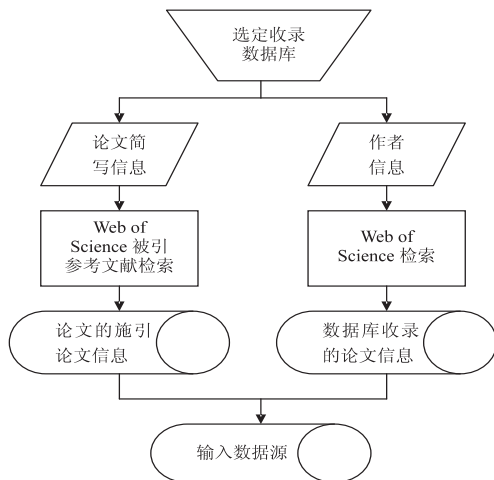


图2 数据获取流程图

通过以上两步工作,将所需要的信息准备到位,最后将该文本文件命名为要统计的作者的英文标准形式(统计的作者肯定是论文作者之一)。程序读取该文件的信息,生成排除自引的统计结果。

2.2 开发环境与平台 软件使用Java语言开发,JDK版本为1.6,编程工具是NetBeans6.7,操作系统为Windows7。应用环境为Java虚拟机1.6版本以上即可。

2.3 关键问题分析

2.3.1 论文简写信息与收录论文信息的匹配。被引论文的作者可能是多个,要排除自引就必须找到被引论文的全部作者,再根据施引论文和被引论文的全部作者找出其中是否有相同的作者。论文简写信息,如前文提到的“ZHAN XW J AM CHEM SOC 2007 129 7246”,其中前两个单词为第一作者的名称,中间的“J AM CHEM SOC”为期刊的简写信息,剩余的数

字为年卷页数等信息,需要根据这些信息在收录论文信息中找到匹配的论文详细信息。

第一步,判断第一作者,将论文简写信息中的前两个词与收录论文信息中作者信息的第一个作者进行匹配。如果第一作者匹配成功,则进入下一步判断。

第二步,根据 ISSN 号匹配。ISSN 表是 Excel 表,表中有两列数据,第一列为期刊简称,第二列为 ISSN 号。选择论文简写信息中第一作者后数字前面的信息,如“J AM CHEM SOC”,经过查找该期刊的 ISSN 号为 0002-7863。用该 ISSN 号与收录论文信息中的 ISSN 号匹配。如果匹配,则说明可能是同一篇文章,反之则不是同一篇。因为 ISSN 表可能没有包含全部期刊的 ISSN 信息,因此会出现期刊找不到对应 ISSN 号的现象,针对这种情况,将期刊信息与后面的年卷页信息一起进入下一个判断步骤,如例子中的“J AM CHEM SOC”和“2007 129 7246”一起进入下一步。基于灵活的考虑,程序将 ISSN 号匹配设为可选项,如果选择 ISSN 匹配,则选择对应的 ISSN 表;如果不选,则直接进入下一步的判断过程。

第三步,判断论文简写信息中剩余的其他信息,如果 ISSN 匹配成功则剩余信息为年卷页信息,如果没有找到期刊 ISSN 号则为期刊简写和年卷页信息。将这些剩余信息与收录论文的来源信息(Source)一一匹配。这里引入容错值的概念,即为剩余的其他信息在收录论文的来源(Source)中不存在的个数。首先在所有收录论文中查找是否存在完全匹配,即错误数为 0 的收录论文,如果有则匹配成功,如果没有,则判断是否存在错误数为 1 的收录论文,即论文简写信息中剩余的其他信息有一个在收录论文的源信息中不存在,以此类推。程序的容错值可以从 0、1、2 中自由选择,默认选择为 1。

论文简写信息与收录论文信息的匹配流程如图 3 所示。

2.3.2 期刊缩写对应 ISSN 表。前文中在论文简写信息与收录论文信息的匹配判断中,使用到了期刊缩写对应 ISSN 的表格。该表对匹配判断甚至于最终结果产生重大影响,因此需要下载信息正确且期刊信息比较全面的表格。这里笔者使用的是 JCR 期刊缩写表,JCR 作为 ISI 出版的《期刊引用报告》^[5],所有数据包括数据源全部使用 ISI 的数据能够保证其一致性。ISSN 表包含两列,第一列为期刊缩写,第二列为对应的 ISSN 号。该表的信息可以在保证信息正确的前提下由工作人员手动补充。

2.3.3 对作者名称匹配的处理。由于英文调用中文名称没有统一的规则,同一位作者可能会出现多种格式,针对多种格式,需要找出一种比较合理的方

法,匹配作者的名称。

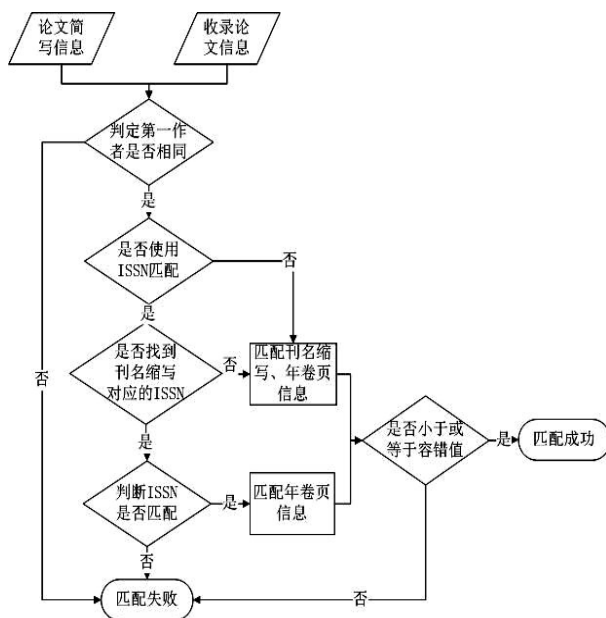


图3 论文简写信息与收录论文信息的匹配

以本文作者名称为例,可能出现的英文名称有“Ma, HS”“Ma, HS”(中间有空格)“Ma, HS(Haishou)”“Ma, HS(Haishou)”(中间有空格)“Ma, HS(Hai-shou)”等多种情况。名称的多样化给自动识别工作带来了很大的不便。综合各种因素考虑,决定采用忽略括号(包括括号内的内容)、逗号、空格、英文句号等特殊符号,将姓首字母大写其余字母小写,名首字母大写的方式,标准如“MaHS”。这样做的优势主要有:a.使问题简单化,不用考虑多种情况,去掉特殊符号可以提高找全的概率。b.大小写区分是为了防止特殊情况下发生错误。如李楠楠,英文为 LiNN,如果大小写不区分李楠楠、林楠都可以写成 linn,尽量防止错误的发生。c.中文名称写成英文如 MaHS 可能不止一人,因此根据字母不能完全判断准确。为了便于人工排查,软件记录找到的作者名称并提示用户,用户只需根据找出的作者名人工判断是否为同一作者。

在作者名称匹配过程中,优先考虑自引排除的召回率,正确率可以在后期人工判断过程中得到保证,软件结合人工判断,大大提高了排查自引的正确性和效率。

3 应用效果分析

3.1 界面设计 为方便引证统计软件的使用,使用 Netbeans 图形化界面工具完成软件的 UI 界面,并添加了参数的设置功能。UI 界面可以使软件的操作更加舒适、简单,软件的功能一目了然。图形界面的功能包括是否使用 ISSN 表选项,ISSN 表路径选择,数据源文件的路径选择,功能模块的选择,容错值的选择,以及分析文件生成结果后打开结果文件。具体的 UI 界面

如图 4 所示。



图 4 软件 UI 界面设计

3.2 结果输出 软件运行结果输出到 Excel 表格中。结果主要由两部分组成:引用次数统计和引用文献目录。

3.2.1 引用次数统计。该部分主要展示引用次数统计,由序号、第一作者、出处、总引次数、他引次数和自引文章列表组成。其中的自引文章列表由自引文章的序号组成,如果该论文简写信息没有找到对应的收录论文信息,则在自引文章列表的最后会显示“无收录”的字样。效果图如图 5 所示。

3.2.2 引用文献目录。该部分结果展示的内容主要由数据源的原始记录和辅助信息组成。原始记录是从第一条论文简写信息开始的,而不是从文件开头的收录论文信息开始。在论文简写信息的右边一格为在 ISSN 表中查找到的期刊的 ISSN 号。原始记录的作者信息对应的右边表格的信息分别为自引的作者名称、对应收录论文的作者列表、对应收录论文的序号、对应收录论文的来源(即 Source)。这样可以很方便地查看论文简写信息是否与收录论文信息有正确的对应关系,以及确认自引的作者名称是否同时出现在施引论文和被引论文的作者名单中。其中论文简写信息、ISSN 号和自引作者名称都以加粗字体显示。效果图如图 6 所示。

3.3 效果分析 使用该引证统计软件对工作人员先前人工统计的多组数据进行测试,通过对两者结果的对比分析发现:a. 利用软件进行引证检索引用统计分析,特别是针对批量论文检索后的数据处理,统计一个文件时间不到 1 秒,极大提高了工作效率;b. 软件统计出的结果比较满意,其中论文简写信息与收录论文信息匹配正确率接近 100%,施引论文判断正确率在

95% 以上;c. 可以进行灵活的人工干预,软件特殊的结果展现形式使得工作人员非常便利地对统计结果进行判断和更正。

表 1 合作者测试实验数据分析

论文简写 信息数	论文简写信息与 收录论文信息		论文简写信息 与收录论文信息		施引论 文数目	软件判断 正确条目	软件判断 正确率
	匹配正确数		匹配正确率				
20	20		100%		333	320	96.1
108	107		99.1		1491	1425	95.6
10	10		100		318	311	97.8

该软件目前已经在中国科学院国家科学图书馆查新检索中心广泛使用,同时也推广至中国科学院图书情报系统使用。

	A	B	C	D	E	F
1	序号	第一作者	出处	总引次数	他引次数	自引文章列表
2		ZHU L	HOLOCENE 2009 18 831			
3		ZHU LP	HOLOCENE IN PRESS 2008 18			
4		1 ZHU LP	HOLOCENE 2008 18 831	17	9	2 3 4 6 10 11 13 15
5		TIAN L	J GEOPHYS RES 2006 111			
6		TIAN LD	J GEOPHYS RES 2006 111			
7		2 TIAN LD	J GEOPHYS RES-ATMOS 2006 111 ARTN D13103	20	9	2 5 7 9 10 11 13 14 16 19 20
8		MA Y	J GEOPHYS RES 2006			
9		3 MA YM	J GEOPHYS RES-ATMOS 2006 111 ARTN D10305	15	7	1 4 5 6 9 10 11 12
10		4 XU BQ	P NATL ACAD SCI USA 2009 106 22114	5	4	5
11		5 YAO TD	GLOBAL BIOGEOCHEM CY 2008 22 ARTN GB4017	4	2	3 4 无收录
12		DUAN K	GEOPHYS RES LETT 2007 34			
13		DUAN KQ	GEOPHYS RES LETT 2007 34 L1810			
14		6 DUAN KQ	GEOPHYS RES LETT 2007 34 ARTN L01810	15	10	3 8 12 14 15
15		7 LUO TX	ECOGRAPHY 2009 32 526	1	0	1
16		8 XU B	P NATL ACAD SCI USA 2008 105 2211	2	2	
17		9 MAY PW	J PHYS-CONDENS MAT 2009 21 ARTN 364203	7	5	2 5
18						

图 5 引用次数统计结果示意图

	A	B	C	D	E
1	原始记录	自引作者	收录文章Authors	收录文章序号	收录文章Source
2	ZHU L HOLOCENE 2009 18 831	0959-6836			
3	ZHU LP HOLOCENE IN PRESS 2008 18				
4	ZHU LP HOLOCENE 2008 18 831	0959-6836			
5					
6	Record 1 of 17				
7	Author(s): Mischke, S (Mischke, Steffen); Zhang, CJ (Zhang, Chengjun)	Author(s): Zhu, L; 1		7	Source: HOLOCENE, 18 (5): 831-839 AUG 2008
8	Title: Holocene cold events on the Tibetan Plateau				
9	Source: GLOBAL AND PLANETARY CHANGE, 72 (3): 155-163 JUN 2010				
10	ISSN: 0921-8181				
11					
12	Record 2 of 17				
13	Author(s): Daut, G (Daut, G.); Mausbacher, R (Zhu, L (Zhu, L))	Author(s): Zhu, L; 1		7	Source: HOLOCENE, 18 (5): 831-839 AUG 2008
14	Title: Late Quaternary hydrological changes inferred from lake level fluctuations of Nam Co (Tibetan Plateau, China)				
15	Source: QUATERNARY INTERNATIONAL, 218 (1-2): 86-93 MAY 1 2010				
16	ISSN: 1040-6182				
17					
18	Record 3 of 17				
19	Author(s): Schutt, B (Schuett, Brigitta); Berking, Schwalb, A (Schwalb)	Author(s): Zhu, L; 1		7	Source: HOLOCENE, 18 (5): 831-839 AUG 2008
20	Title: Late Quaternary transition from lacustrine to a fluvo-lacustrine environment in the north-western Nam Co, Tibetan Plateau, China				
21	Source: QUATERNARY INTERNATIONAL, 218 (1-2): 104-117 MAY 1 2010				
22	ISSN: 1040-6182				
23					
24	Record 4 of 17				
25	Author(s): Zhu, LP (Zhu, Liping); Peng, P (Peng, Zhu, LP (Zhu, Liping))	Author(s): Zhu, L; 1		7	Source: HOLOCENE, 18 (5): 831-839 AUG 2008
26	Title: Ostracod-based environmental reconstruction over the last 8,400 years of Nam Co Lake on the Tibetan plateau				
27	Source: HYDROBIOLOGIA, 648 (1): 157-174 JUL 2010				
28	ISSN: 0018-8158				

图 6 引用文献目录结果示意图

4 结 语

本文从工作中的实际需求出发,完成了一个基于 ISI Web of Knowledge 平台检索结果的引证统计分析软件,完成了自引和他引的自动分析和统计,并将结果以特定的格式输出到 Excel 表中。按照要求下载数据,即可使用该软件。经过测试,在保证正确率的前提下,极大地提高了工作效率。目前该软件还存在一定 (下转第 135 页)

.....

(上接第 152 页)

的不足,如中文作者名称的引用格式不统一,期刊 ISSN 号可能随时间变化或出现多个 ISSN,笔者会在以后的工作中不断对其进行改进。

参 考 文 献

[1] 金碧辉,刘雅娟. 期刊评价与影响因子、被引频次[J]. 中国科技期刊研究, 1998,9(4):239-241

[2] ISI Web of Knowledge[EB/OL]. [2011-07-02]. <http://apps.webofknowledge.com>

[3] Web of Science General Search[EB/OL]. [2011-08-01]. http://apps.webofknowledge.com/WOS_GeneralSearch_input.do?product=WOS&search_mode=GeneralSearch

[4] Web of Science Cited Reference Search[EB/OL]. [2011-08-10]. http://apps.webofknowledge.com/WOS_CitedReferenceSearch_input.do?product=WOS&search_mode=CitedReferenceSearch

[5] Journal Citation Reports [EB/OL]. [2011-06-20]. <http://admin-apps.webofknowledge.com/JCR/JCR>

(责编:刘影梅)