论文收录证明辅助系统的设计与应用

□ 孔云 资芸 杨婷 薛秀珍 / 昆明理工大学图书馆 昆明 650093

摘要：出具论文收录证明是高校图书馆信息咨询部的重要业务之一，其基本流程为客户提出申请，图书馆员检索引文数据库，从检索结果文件提取论文信息，生成报告等。其中最耗时的环节为从引文文件提取信息的蛮力过程，所耗时间和论文篇数呈线性增长。文章首先分析了开具检索证明的业务流程和业内研究现状，其次分析了检索结果文件，接着设计和开发了论文收录证明报告辅助系统，最后以系统在本校超过三年的使用效果说明：该系统极大地提高了出具收录证明的速度和信息咨询部的工作效率，是一套具有参考和推广价值的系统。

关键词：论文收录证明，自动化，信息咨询，信息服务
DOI：10.3772/j.issn.1673—2286.2013.09.008

1 引言

1.1 论文收录证明服务介绍

根据我国的国情和相关部门的规定，高校教师或其他科研机构的科研人员在申报国家及省部级各类奖项、课题、科技成果鉴定验收、科研成果奖励、个人职称评定等工作时，申报人员提供的论文须由相关部门审定，其中以独撰或第一作者发表的论文被SCI、EI等收录的，须由具有国家认可的资质单位开具收录或引用检索报告。因此国内具备资质的高校图书馆为申报人员开展了论文收录证明服务，一般由图书馆的信息咨询服务部承担此项服务，以下简称为信咨部。

1.2 引文数据库平台简介

ISI Web of Knowledge简介[1]：此平台以三大引文数据库SCI、SSCI、A&HCI为核心，同时还有两个化学信息事实型数据库CCR、IC和三个引文数据库CPCI-SSH、SCIE、CPCI-S(ISTP)。兼具知识的检索、提取、管理、分析与评价等多项功能。

Engineering Village简介[2]：该平台是最权威的工程、应用科学领域文献检索平台。它提供最专业、内容最丰富的工程科学数据库和相应的科技文献检索，以及全球优秀工程科学期刊的全文在线访问服务，提供著名的工程索引EI功能。

2 业内研究现状和存在的问题

2.1 业内相关研究现状

以万方和CNKI数据库为依据，用“检索证明”、“检索报告”、“代查代检”、“自动化管理”、“网络化管理”、“计算机管理”以及“服务平台”等为检索词，分别在主题、关键词和摘要中进行检索，发现相关的研究成果和系统可以归纳为[3,4]：（1）进行查新项目档案管理；（2）进行量化管理或绩效管理；（3）建立网上服务方式方便用户；（4）进行质量控制，通过对业务流程的控制，提升服务质量；（5）建立知识库。

从已有的文献来看[4-10]，目前图书馆界针对论文收录证明自动化系统的研究几乎为空白。

2.2 当前出具检索证明的弊端

出具论文收录证明的基本流程为：客户填写检索申请表，图书馆员根据申请表选择引文数据库，检索客户需要的论文，筛选论文，经客户同意后输出记录文件，图书馆员分析文件内容，按格式生成检索报告文档，盖章签字、支付服务费完成服务。其服务流程如图1所示。

在和信咨部的专家反复沟通业务过程后，一致认为：论文收录证明业务流程最耗时的环节为从分析文件内容到按格式生成检索报告文档的过程，其所耗时间和客户要求检索的论文篇数成正比例关系。通常来说，提取一篇论文信息所耗费的时间为30分钟左右。如果一个

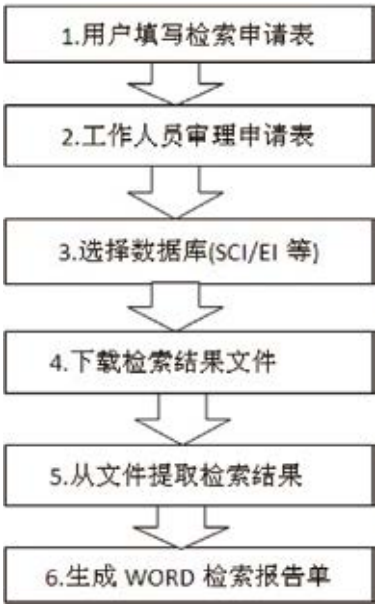

图1 论文收录证明报告流程

图2 自动解析引文文件流程图

客户一次要求检索的论文篇数为10篇（这种情况在我们学校比较普遍），则所耗费时间为30*10=300分钟。实际情况要多于这个时间，因为工作人员不是机器，越往后就越疲劳，而且中间环节还会被其他业务中断。据信咨部反映，提取信息的过程，基本是一个寻找信息、组合信息、复制、粘贴到Word的过程。信咨部希望能够把这个过程自动化，以提高他们的工作效率。

3 论文收录证明辅助系统的分析与设计

作者和信咨部的专家沟通后，明确了出具论文收录证明的流程(见2.2节介绍)。国内需出具收录证明的引文数据库已经被整合到ISI Web of Knowledge（为了便于讨论，以SCI为简称）和Engineering Village（为了便于讨论，以EI为简称）两大引文检索平台，这一工作为本文想实现的辅助系统提供了有限的数据来源，系统只需要分析两种数据格式：即SCI和EI引文数据格式，有效降低了系统实现的难度和复杂度。

3.1 论文收录证明辅助系统的设计思路

要实现论文收录证明的全部自动化需要检索平台提供功能完整和灵活的API便于第三方开发者调用。从目前掌握的信息来看，Thomson Reuters公司于2012年2月开放了SCI的一个Web服务[11]：通过此API，机构用户可以实时查询和获取该机构的元数据信息，包括作者，文章标题，Source数据，关键字和文章唯一标识号。此API主要是为方便学术机构从其主页或机构知识库接入SCI平台。针对出具论文收录证明，此API至少有两方面的不足：首先是返回的数据有限，只返回5个字段；其次，只返回本机构的数据。论文收录证明报告要求返回较完整的数据，便于适应不同的报告模板；其次客户的范围是广泛的：包括不同高校、不同的科研机构，而不是限制在一个机构内。至于EI，目前还没有提供开放的接口。因此，以目前的情况看，通过API的方式是不可行的。

从2.2节的讨论可以看出，出具证明的瓶颈在于：从引文文件到生成检索报告的过程，几乎占去了整个过程90%的时间。如果可以解决此瓶颈，将极大提高出具证明的效率。因此本文设计和开发了论文收录证明辅助系统：主要是解决从分析文件内容到生成检索报告的自动化问题。首先由图书馆员在两大检索平台上检索到客户的论文并下载引文文件数据，然后使用辅助系统自动生成检索报告：图书馆员上传引文结果文件到辅助系统，系统按照算法自动解析文件，然后生成并返回网页形式的检索报告，检查无误后，自动生成Word格式的正式报告。其流程如图2所示。

3.2 引文文件分析

3.2.1 SCI引文文件分析

在获取SCI引文库检索结果后，按如下步骤输出检索结果文件：（1）选择全记录方式，目的是获取论文的相关信息，为后续解析文件提供完整的信息；（2）选择保存文件的方式为制表符分隔的格式(Win,UTF-8)，这里规定字符编码为UTF-8，目的是为避免编码出现乱码。分析导出文件，可以看到SCI引文为论文提供了58个字段，提供的内容是以二维表的方式呈现的，这为计算机程序自动分析和提取内容提供了便利。SCI引文文件格式如表1所示。

3.2.2 EI引文文件分析

在获取EI引文库检索结果后，

选择下载选中的文章，在下载页面，选择记录详情（record detail）和下载格式（plain text format ASCII）。分析下载的文件，可以看到EI引文的文件格式是以<record + 编号>，换行，字段名称+“:”+字段内容+换行的方式陈列，第一篇文章的内容显示完后，换行，然后又是以<record + 编号>，换行，字段名称+“:”+字段内容+换行的方式显示内容。EI引文为论文提供了30个左右的字段信息，抽象后的文件格式如表2所示。

3.3 检索结果算法设计

3.2节分析了SCI和EI两大引文结果文件结构，为设计计算机算法提供了基础。

3.3.1 SCI算法设计

由3.2节的分析可知，SCI引文文件的内容为一张二维表，表头为每篇论文的字段名称，每篇文章对应二维表的一行，每行提供58列（即58个属性），二维表的行数由检索到的论文篇数确定。为了便于计算机程序操作，定义如下数据结构：

（1）SCI对象SCI(PT,AU,BA,BE,GP,AF,BF,CA,TI,SO,SE,BS,LA,DT,CT,CY,CL,SP,HO,DE,ID,AB,C1,RP,EM,RI,FU,FX,CR,NR,TC,Z9,PU,PI,PA,SN,BN,J9,JI,PD,PY,VL,IS,PN,SU,SI,MA,BP,EP,AR,DI,D2,PG,P2,WC,SC,GA,UT)，其中SCI为对象名称，是每篇论文字段信息的集合，括号内的58个字段为SCI引文为每篇论文提供的字段名称。

（2）SCI对象数组

为了存储所有的SCI对象，定义对象数组List<SCI>={SCI1,SCI2,…,SCIn}。List<SCI>是一个线性表，线性表的元素为SCI对象。SCI引文文件解析流程如图3所示。

表1 SCI引文文件格式

字 段	AU	AF	TI	…
内容	作者	作者全名	论文标题	省略字段

表2 EI引文文件格式

<RECORD 1>		
Accession number	:	20113314234118
Title	:	Improving hadoo.
…	:	…
<RECORD 2>		
Title	:	Apache hadoop ..
…	:	…
<RECORD N>		

（3）SCI引文算法sciAnalyze关键代码如下所示：

```
1. 输入：sciBufferedReader //输入为SCI引文检索结果文件
2. 输出：sciList //返回SCI对象数组
3. List<Sci> sciList = new ArrayList<Sci>(); //新建SCI数组
4. String s = null //定义字符串变量
5. int i = 0 //定义标志符
6. while ( ( s = sciBufferedReader.readLine())!= null) {
        if( i == 0 ) {
                ++i;
                continue;
                }// 过滤引文表头，不必存储
7. String sp[] = s.split( "\\t" );//根据水平制表符分隔字符串
8. SCI sci = new SCI();//新建SCI对象
9. sci.setPT( sp[ 0 ] ); //为对象属性赋值
10. sci.setAU( sp[ 1 ] ); //为对象属性赋值
        ...
11. sci.setUT( sp[ 57 ] ); //为对象属性赋值
12. sciList.add( sci ); //添加SCI对象到对象数组
    }
13. sciBufferedReader.close();//关闭引文文件输入流
14. Return sciList; //返回SCI对象数组
```

3.3.2 EI算法设计

由前面的分析可知，EI引文文件的内容如表2所示。为了便于操

作，定义如下数据结构：

（1）EI对象EI(accessionNumbertitle,authors,authorAffiliation,correspondingAuthor,souceTitle,abbreviatedSourceTitle,volume,issue,monographTitle,issueDate,publicationYear,pages,articleNumber,language,issn,eissn,isbn,isbn10,documentType,conferenceName,conferenceDate,conferenceLocation,conferenceCode,sponsor,publisher,abstract,numberOfReferences,mainHeading,controlledTerms,uncontrolledTerms,classificationCode,doi,database)，其中EI为对象名称，是EI引文所能提供的所有字段信息的集合，括号内的字段为EI引文为每篇论文提供的字段名称。

（2）EI对象数组

为了存储所有的EI对象，定义对象数组List<EI>={EI1,EI2,…,EIn}。List<EI>是一个线性表，其元素为EI对象。EI引文文件解析流程如图4所示。

（3）EI引文算法eiAnalyze关键代码如下所示：

```
1. 输入：eiBufferedReader //输入EI引文检索结果文件
2. 输出：eiList//返回EI对象数组
3. List<Ei> eiList = new ArrayList<Ei>();//初始化EI对象数组
4. String s = null; //定义字符串变量
5. Ei ei = null; //定义EI对象
6. While((s = eiBufferedReader.readLine()) != null){ //按行读取文件
7. if( s == null || s.equals( "" ) || s.contains( "<RECORD" )
8. || s.contains( "Compilation and indexing terms" )){
9. continue;
}//读到空行或者对象结束标记则跳转
10. String sp[] = s.split( ":" ); //以冒号为标记分离字符串
11. if( sp[ 0 ].equals( "Accession number" )){
12. ei = new Ei();//读到对象的第一个属性时，新建对象
13. ei.setAccessionNumber( sp[ 1 ] ); //对象属性赋值
14. } else if( sp[ 0 ].equals( "Title")){
ei.setTitle( sp[ 1 ] ); //对象属性赋值
15 } else if( sp[ 0 ].equals( "Authors" )){
16. ei.setAuthors( sp[ 1 ] ); //对象属性赋值
17. } else if( sp[ 0 ].equals( "…" )){
18. …; //省略对象属性赋值
} else if( sp[ 0 ].equals( "Database" ) ) {
ei.setDatabase( sp[ 1 ] );
eiList.add( ei ); //读到对象的最后一个属性时，加入到对象数组中
ei = null;
}
}
19. eiBufferedReader.close();//关闭文件输入流
20. Return eiList; //返回对象数组
```

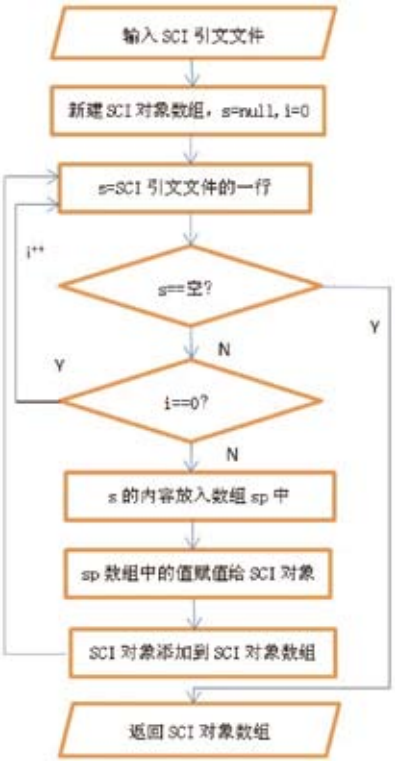

图3 SCI引文文件解析流程

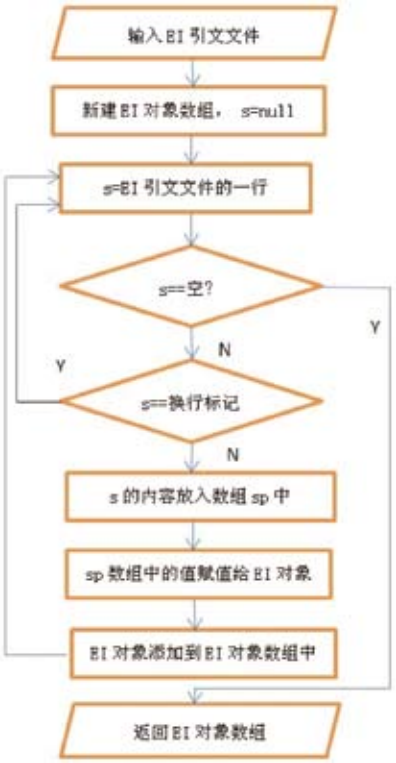

图4 EI引文文件解析流程

3.4 辅助系统完整的算法

（1）图书馆员上传引文检索结果文件；

（2）系统选择解析算法：sciAnalyze()或eiAnalyze()；

（3）生成检索报告，并按显示格式返回网页形式的检索结果；

（4）检索结果自动导入到Word文档；

（5）检查和调整Word文档，形成正式检索报告；系统流程如图2所示。

4 系统实现和应用效果

4.1 系统采用J2EE平台

J2EE平台具有开发结构简单、开发效率高、移植性强、重用性好、易于维护、伸缩性强、被广泛接受等优势，是企业级应用系统事实上的标准。对于信息技术日新月异的时代，考虑系统的可扩展性，是企业应用的首选开发平台。

4.2 系统部署平台

论文收录证明自动生成系统的部署环境为：Intel(R) Xeon(R) CPU E5420,主频为2.50GHz，RAM 1.0 GB，Windows Server 2003 Enterprise Edition SP2，JDK1.6.-0.26，Web容器采用开源且性能稳定的Apache-Tomcat-6.0.32[12]，本系统在开发过程中综合使用了HTML、JAVA SCRIPT、VELOCITY[13]、JAVA和开源文件上传组件commons-fileupload-1.2.1.jar[14]等技术。

4.3 系统在我校图书馆的应用效果

2010年12月，我校被批准为“教育部部级科技查新工作站筹建单位”。论文收录证明服务是我校图书馆的重要业务之一，自系统应用以来，为我校高端人才引进、创新团队建设、人才培养、重点实验室建设、重点学科与专业建设、专业评估、专业技术职称评审、教师绩效考核等工作和校外其他单位开展的论文收录与检索证明服务提供了有力的支持。2009、2010、2011年完成论文收录证明报告分别为112项、228项、391项，2012年截止到11月12日，已完成收录证明报告650项。历年累计完成SCI检索3577篇，EI检索4062篇，ISTP检索377篇，共累计完成8016篇检索证明服务。

自本系统应用以来，信咨部做论文收录证明报告的速度有了显著的提高，不但显著降低了工作辛劳度，而且可以把更多的时间投入到其他信息咨询服务中去；同时极大地缩短了客户开具检索证明的等待周期，为客户节约了宝贵的时间。图5所示：为检索文件输入入口，图书馆员根据引文选择文献类型，然后上传文件到文件自动解析系统，服务器将自动生成检索报告，并返回网页形式的检索结果，如图6所示，为EI类型的检索结果，从输入文件到生成固定格式的检索报告所用的时间在秒级以内。点击图6左上角的导入Word按钮，程序将自动调用Word组件，并按配置参数生成Word文档，图书馆员只用稍加修饰就可以形成一份检索报告，极大地提高了工作效率。

5 总结与展望

本文首先研究了引文证明业务流程和业内研究现状；然后设计并实现了论文收录证明辅助系统；最后以系统在我校的使用效果说明系统有效提高了信咨部的工作效率，尤其是当同一个作者或科研团队要出具多篇文章的检索证明时，系统所花的时间几乎没增加，让原本是一件痛苦的事情变得十分简便。展望未来，笔者希望ISI Web of Knowledge和Engineering Village引文检索平台能够提供完整和灵活的API服务，让第三方开发机构可以调用接口，从而简化检索过程和自主定制检索报告，进一步提高系统自动化的程度，同时由于每个学校出具检索证明的模板各异，系统下一步将增加后台模板定制功能。

图5 文件自动解析入口

导入Word

检索报告附件（编号：2012-00***）

工大学　　学院　　委托，检索　　年间，发表论文被SCI、EI、ISTP收录的情况，经Web of Science、Engineering Village数

序号	SCI/ EI/ISTP存取号	论文题目	出版物名称	年，卷（期）	页码	出版/会议召开时间	文献类型	作者	SCI/EI/ISTP收录	单位
1	EI: 20113314234116	Improving hadoop performance in handling small files	Communications in Computer and Information Science	2011, 193 CCIS (PART 4)	187-194	July 22, 2011 - July 24, 2011	Conference article (CA)	Mohandas, Neethu (1); Thampi, Sabu M. (1)	EI（核心版）	(1) Rajagiri School of Engineering and Technology, Cochin, India Mohandas, N.
2	EI: 20122115035017	Apache hadoop performance-tuning methodologies and best practices	ICPE'12 - Proceedings of the 3rd Joint WOSP/SIPEW International Conference on Performance	2012	241-242	April 22, 2012 - April 25, 2012	Conference article (CA)	Joshi, Shrinivas B. (1)	EI（核心版）	(1) Advanced Micro Devices, Inc., 7171 Southwest Pkwy, Austin, TX 78735, United States Joshi, S.B. (shrinivas.joshi@amd.com)

图6 检索结果实例

参考文献

[1] 百度百科.ISI Web of Knowledge [OL]. [2013-03-20]. http://baike.baidu.com/view/878678.htm.
[2] 百度百科. Engineering Village介绍[OL]. [2013-03-20]. http://baike.baidu.com/view/1466057.htm.
[3] 郭丽君.高校图书馆科技查新服务调查与分析[J].情报杂志,2012,31(1).
[4] 孙海刚.个性化服务在数字图书馆科技查新中的研究与应用[D].中南大学,2007.
[5] 张天俊.Php&Mysql技术在高校图书馆“代查代检”服务系统开发中的应用[J].情报科学,2003,21(7).
[6] 战玉华,等.代检代查服务系统的开发及应用[J].图书情报工作,2005,49(11).
[7] 郑菲,等.中国科学院科技查新检索服务平台的设计与实践应用[J].现代图书情报技术,2010(11).
[8] 马骅,等.多校区环境下科技查新:以南京大学图书馆为例[J].图书馆学研究(理论版),2010(2).
[9] 马景娣,等.基于J2EE的科技查新综合信息系统的设计与实现[J].现代图书情报技术,2004(8).
[10] 但旺等.科技查新业务管理系统设计分析[J]. 图书馆学研究,2008(4).
[11] Thomson Reuters. Web of Science [OL]. [2013-03-20]. http://wokinfo.com/products_tools/products/related/webservices/.
[12] The Apache Software Foundation. Apache Tomcat [OL]. [2013-03-20]. http://tomcat.apache.org/index.html.
[13] The Apache Software Foundation. The Apache Velocity Project [OL]. (2010-11-29) [2013-03-20]. http://velocity.apache.org/.
[14] The Apache Software Foundation. Commons FileUpload [OL]. (2010-07-30) [2013-03-20]. http://commons.apache.org/fileupload/index.html.

作者简介

孔云（1982-），馆员，研究方向：图书馆自动化。E-mail: 920581344@qq.com
资芸（1973-），副研究馆员，研究方向：数字图书馆。

Design and Application of Assistant System on Paper Published Proof

Kong Yun, Zi Yun, Yang Ting, Xue Xiuzhen / Lib of Kunming University of Science and Technology, Kunming, 650093

Abstract: Showing paper published proof is one of the most important business for the information department in the university library, whose basic process includes customer submitting an application, librarian retrieving database, extracting information from the result, generating a report, etc. The most time-consuming part of this process is to extract information from the downloaded result, which is a procedure of brute force. The time consuming is growing linearly as the number of papers increases. This article first analyses the business process and survey the background on showing paper published proof, followed by analyzing the search result, and then designing and developing an assistant system about paper published proof, at last the system's application effect for more than three years in our library demonstrates that the system greatly improves the librarian's working efficiency and accelerates the speed on making a report. It is really a system of promotional and reference value on the industry.

Keywords: Paper published proof, Automation, Information consultation, Information service

（收稿日期：2013-04-14）