一种领域专家文献自动收集系统①

廖晓锋[1,2], 王永吉[2,3], 周津慧[2], 关 贝[2]

[1](南昌大学信息工程学院 信息管理系，南昌 330029)
[2](中国科学院 软件研究所 基础软件国家工程中心，北京 100190)
[3](中国科学院 软件研究所 计算机科学国家重点实验室，北京 100190)

摘 要：设计并实现了一种自动专家文献信息收集系统（BibCollector）。收录对象针对计算机科学技术领域的专家学者，收集范围涵盖国内外主要的全文数据库(SpringerLink, IEEE Xplore,ACM Digital Library, Elsevier Science Direct,中国知网 CNKI 和万方数据)和常用的引文数据库(SCI,EI,ISTP,CSCD)及专利数据库(Derwent)。该系统使用专家姓名和工作单位作为标识,判断记录相关性和去除重复项,生成的文献列表具有较高的准确度。该系统同时收集专家所发表的中文和外文文献，因此无论相比国外和国内的类似系统，该系统都具有数据来源更丰富的优势。该系统能为相关的文献收集工作节省大量人力。

关键词：专家文献；文献自动收集；BibCollector；重复检测

Automatic Bibliography Integration System for Domain Experts

LIAO Xiao-Feng[1,2], WANG Yong-Ji[2,3], ZHOU Jin-Hui[2], GUAN Bei[2]

[1](Information Engineer School, Nanchang University, Nanchang 330029, China)
[2](National Engineering Research Center for Fundamental Software, Institute of Software, The Chinese Academy of Sciences, Beijing 100190, China)
[3](State Key Laboratory of Computer Science, Institute of Software, The Chinese Academy of Sciences, Beijing 100190, China)

Abstract: We designed and implemented a system called BibCollector which can automatically collect the bibliography information from different databases. This system is targeted at experts in Information Technology (IT) domain. The databases covered include the most used ones such as SpringerLink, IEEE Xplore, ACM Digital Library, Elsevier ScienceDirect. Two main Chinese databases CNKI and Wanfang are also included. The citation databases that are covered include: Science Citation Index, EI, ISTP, CSCD. Besides these, the Derwent patent database is also included. We presented a method by using the name and affiliation/address of a person to accurately query from these databases. We also developed some algorithms to exclude the unrelated records and identify the duplicate ones. Comparing to the overseas and domestic counterparts, our system has advantages of richer record sources and more accurate results.

Key words: bibliography collection; bibCollector; duplicate identification

1 引言

一个完整而准确的已发表文献集合，是了解专家的最好途径。然而收集并维护特定专家的文献列表需要花费大量的精力。因为文献散布于各种期刊及会议，并且被不同的数据库所收集。数据库厂商的收录范围也有交叉重叠。在这种局面下，为了准确全面的获得专家的文献集合，通常的做法是人工遍访领域相关的所有数据库进行查询，然后将从多数据源收集到的查询结果进行汇总整理，去除重复和不相关的数据。每个数据库都以自定义方式进行检索和提供查询结果，因此用户需要学习各数据的检索方法，还要把从各数据库的检索结果进行整理和汇总，这个工作需要花费

① 收稿时间:2011-10-01;收到修改稿时间:2011-11-04

很多人力[1-3]。

2 相关工作

围绕这个目标，已经有了适当的研究并出现了一些专家信息收集分析平台，如：DBLP，ACM Author Profile, CDBLP，ArnetMiner。这些平台从各种途径获取文献信息，经过分析，形成人物写照，供用户查询[4]。

作为世界上最大的计算机领域文献记录数据库，DBLP(Digital Bibliography & Library Project)[5]是以人物为中心的文献记录数据库。用户只需输入作者姓名即可获得该作者所发表的论文列表。从 1993 年发展至今，已经收集了 120 多万条文献记录，涵盖了计算机领域内绝大部分期刊和会议的论文记录[5,6]。由于数据格式不规范等原因，DBLP 需要适当人工参与来保证文献记录质量，并且有些数据是由手工录入系统[7]。ACM 则于近期推出 Author Profile Page1 服务 β 版，该服务在 ACM 所拥有的数据基础上，为计算机领域内著名学者生成主页，在页面上列出所发表文献信息，并提供文献下载次数、引用次数等统计信息，以及合作者列表。

对于中国学者而言，这两个系统提供的文献列表都无法全面的反映他们的研究成果。因为 DBLP 和 ACM 中收集的文献记录都是英文信息，并不收录中文文献。由于汉字与汉语拼音并非一一对应，所以 DBLP 对中国作者用拼音进行姓名聚类时，不可避免的会产生不精确聚类。ACM 的数据来源仅仅限于其自身收录的文献，而除了 ACM 之外，还有类似的数据库，如 SpringerLink,Xplore(IEEE), ScienceDirect(Elsevier)等，而且这些数据库的收录来源不尽相同，所以 ACM Author Profile 所提供的文献列表只能部分反映学者的研究成果。

ArnetMiner2 是一个专家搜索系统，其数据来源有两部分，一是可公开获取的 DBLP，二是从 Web 抓取。首先定义 Researcher Profile 本体，然后把研究者名字作为关键词，使用搜索引擎（Google）从互联网上获取一个网页列表，对页面进行分类并抽取相关信息[8]。ArnetMiner 的不足之处在于从网页抽取信息，所提供的专家信息不够准确。

如其名字所示，CDBLP(Academic Search In China)3 收录重点是中文文献，其收录范围限于国内 13 种核心期刊及会议。因为科技文献绝大部分以英文撰写，国内学者也普遍在国际期刊及会议发表文章，所以 CDBLP 难以对学者的研究工作给出科学的全貌。

Google Scholar4 也提供“学者检索”服务，用于搜索特定作者的文献，但其检索结果中包含很多不相关记录，准确度不高。

由此可见，虽然有许多工作致力于检索特定学者的全面准确的文献列表，但鉴于信息分布广泛以及形式多样，现有的系统难以满足本文开头所提出的实际问题，即解决图书馆员所面对为特定机构内大量研究人员收集全面准确的文献列表的任务。要想获得特定专家的全面准确的文献信息，最可靠的途径是人工查询多个数据，然后对多个数据源得到的记录进行汇总整理。

为了减轻文献记录收集过程中的人力开销，我们设计并实现了一个自动文献记录收集系统，用于为特定机构内的专家收集全面准确的文献列表。我们的领域设定为计算机领域。为了保证准确性，我们向常用的数据库中提交查询请求来获得文献记录。为了保证文献记录列表的全面性，我们选择的数据库包含了计算机领域常用的外文全文数据库,包括 SpringerLink, Xplore(IEEE),ACM Digital Library, ScienceDirect (Elsevier)；以及常用的中文全文数据库,包括中国知网(CNKI)和万方数据。除了全文数据库之外，还涵盖了常用的引文数据库,包括 SCI,EI,ISTP 和中国科学引文数据库 CSCD(Chinese Science Citation Database)。数据来源中包含了权威中外文电子数据库，使得我们的系统能为中国学者生成全面的文献列表。

文章第 2 节介绍了国内外的相关工作。第 3 节介绍了我们为了达成目标所采取的三条检索策略,以及针对检索结果所进行的数据清洗工作,包括排除不相关记录和合并重复记录的方法。第 4 节介绍了基于这些策略所构建的检索系统 BibCollector 的系统架构。最后在第 5 节对工作进行总结和展望,在此小节中还简短列举了在数据整理过程中发现的各数据库中数据处理的一些不一致现象，以及由此引发的对统一文献记录规范的想法。

3 文献记录集合生成方法

3.1 文献记录检索

目标:全面和准确的收集特定专家的文献记录。

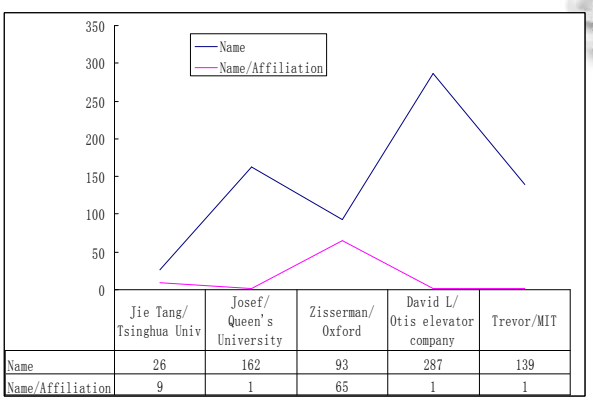
1. 全面指收集得到专家所有的发表文献记录，包括其在不同机构工作期间发表的文献，先决条件是知晓专家姓名及其曾任职的机构名称。

2. 准确指只收集属于该专家的文献记录，排除同名专家的文献记录，并检测出由于被多个数据库收录而导致的重复问题。

我们采用三条策略来实现这两个目标。

方法 1：扩大数据源涵盖范围。为了尽可能全面的获取专家的发表文献记录，我们的数据来源范围涵盖了计算机领域常用的国内外全文数据库和引文数据库，以及专利数据库。包括计算机领域常用的外文全文数据库包括：SpringerLink,Xplore(IEEE), ACM Digital Library, ScienceDirect(Elsevier)；常用的中文全文数据库包括中国知网(CNKI)和万方数据。除了全文数据库之外，还涵盖了引文数据库,包括 SCI,EI,ISTP 和中国科学引文数据库 CSCD(Chinese Science Citation Database)。因为数据来源中同时包含了权威中外文电子数据库，使得我们的系统能为中国学者生成全面的文献列表。

方法 2：使用“作者”和“机构/地址”两种字段进行检索。从数据库中检索特定人物，最直观有效地检索方式就是使用“作者”检索字段，但仅使用“作者”一个字段进行检索，查询结果中会含有很多不相关记录。如果辅以“机构/地址”字段，则能大幅度减少检索结果中的不相关记录数目。图 1 为在 Xplore(IEEE)中，分别使用这两种检索方式进行检索得到的记录数目对比。可以看出，使用“作者+机构/地址”方式能过滤大量不相关记录。



	Jie Tang/ Tsinghua Univ	Josef/ Queen's University	Zisserman/ Oxford	David L/ Otis elevator company	Trevor/MIT
Name	26	162	93	287	139
Name/Affiliation	9	1	65	1	1

图 1 “姓名”和“姓名+地址”两种检索方式对比

表 1 统计了各数据库支持的检索字段以及能提供的文献记录格式。可以看出，“姓名”和“机构/地址”是最通用的检索字段。但也有些数据库不提供“机构/地址”检索字段。对于后面这种情况，则先用“作者”进行检索，然后使用“机构/地址”字段在所获得的查询结果中进行二次检索。SpringerLink，Derwent 数据库属于后面这种情况。除了“姓名”和“机构/地址”字段之外，查询请求中还可以结合其他检索字段对查询请求进行限制，比如“时间”字段。

表 1 常用数据库检索方式及提供的文献记录格式

	数据库	主要检索字段	文献记录格式
1	ACM Digital Library	Title, Author, Keyword, Affiliation	BibTex, EndNote, ACMRef
2	Xplore(IEEE)	Title, Author, Affiliation	BibTex, EndNote, Refworks, ASCII
3	ScienceDirect (Elsevier)	Authors; title; Affiliation;	BibTex, EndNote, Refworks, ASCII
4	ISI Web of Knowledge	标题；作者；地址；	EndNote, Refworks
5	中国科学引文数据库	标题；作者；地址；	EndNote, Refworks
6	中国知网 (CNKI)	题名；关键词；作者；单位；	自定义文本
7	万方	标题；作者；单位；关键词；	XML,HTML,Txt
8	SpringerLink	标题,；作者,	RIS，文本
9	Derwent 专利数据库	主题；标题；发明人；专利号；	
10	DBLP	Authors only; Venues only	BibTex,XML

方法 3：姓名及地址扩展查询。由于数据库中经常使用名字缩写[9,10]，一个作者的名字可能会有多种写法。不同的数据库在名字的处理上采取的方法不一样，比如 IEEE 数据的常用做法是显示作者的全名，而 ISI Web Knowledge 则使用姓加上名首字母的方式。另外 IEEE 中把名字放在姓氏前面，而 ISI Web Knowledge 则是把姓氏放在名前面，而且当名并非单字时，在其中用连接符相连。如下例所示：

1. IEEE ：Da-Zhi Sun; Jin-Peng Huai; Ji-Zhou Sun; Jian-Xin Li; Jia-Wan Zhang; Zhi-Yong Feng;

2. ISI Web Knowledge：Sun DZ (Sun, Da-Zhi) Huai JP (Huai, Jin-Peng), Sun JZ (Sun, Ji-Zhou), Li JX (Li, Jian-Xin), Zhang JW (Zhang, Jia-Wan), Feng ZY (Feng, Zhi-Yong)

由于各数据库对名字的处理方式不统一，导致同

一个作者的名字会有多种不同的表现形式，这为我们的查询带来了挑战。比如一个人的姓名有多种写法，如图 2 所示。也可能多个人拥有相同缩写的情况，如图 3 所示。

为了提高查全率，我们将作者姓名变换成数据库常用的多种缩写形式进行扩展查询，对地址也进行同样的处理。用户提交作者(Name)和机构(Affiliation)后，系统构建出多个查询请求分别提交至不同数据库。

3.2 文献记录清洗

在经过上一步骤从多个数据库中检索得到相关的文献记录之后，接下来要对检索结果进行数据清洗工作，主要完成两个任务：一是排除不相关记录，二是合并重复记录[11,12]。

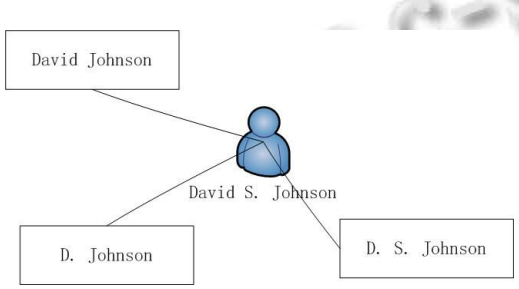


图 2　作者多名

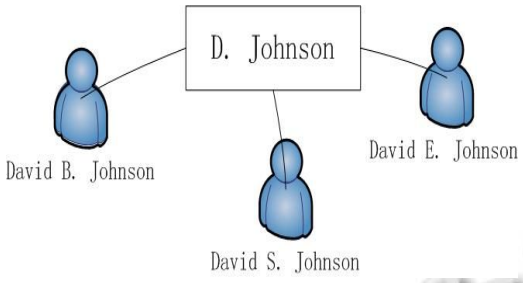


图 3　多人同名

3.2.1 不相关记录排除

在检索过程中，即使使用精确匹配方式进行检索，也会带来不相关记录。下面例举了使用一个常用的中国名“Jing Liu”进行精确匹配查询，表达式为 au:("Jing Liu")，从 IEEE，Springer，ACM 得到的不相关结果。结果中以粗体标识部分与检索要求不符。

对于这种情况，我们通过把查询表达式中的“作者”和“机构/地址”字段与查询结果中的相应字段进行比对来排除。做法如图 4 所示。如果有两个同名的人在同一个机构工作，并且都以机构名义发表了论文。对于这种小概率事件，需要人的参与来进行数据的验证。我们使用 Wiki 作为前端显示，让专家作为用户参与数据的修订。

表 2　以“Jing Liu”作为检索词从 Springer, IEEE, ACM 检获的不相关记录

Springer	*Quantum Wave Equation of Non-conservative System* *Xiang-Yao Wu, Bai-Jun Zhang, Hai-Bo Li, Xiao-**Jing Liu** and Jing-Wu Li, et al.* *International Journal of Theoretical Physics, 2009, Volume 48, Number 7, Pages 2027-2035*
IEEE	*Research on real-time remote video monitoring system design* *Liu Yuejun; Su **Jing; Liu** Feng; Computer and Automation Engineering (ICCAE), 2010 The 2nd International Conference on: 4　2010 : 484 - 487*
ACM	*Personal information management with SEMEX* *Yuhan Cai, Xin Luna Dong, Alon Halevy, **Jing Michelle Liu**, Jayant Madhavan June 2005 SIGMOD '05: Proceedings of the 2005 ACM SIGMOD international conference on Management of data*

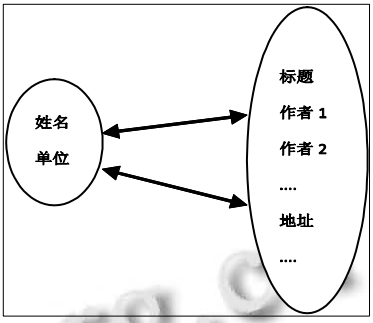


图 4　对比查询表达式和查询结果相应字段

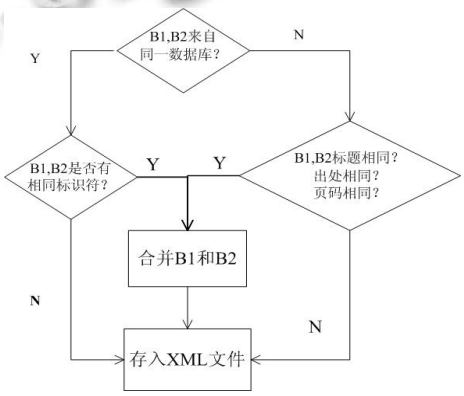


图 5　去除重复记录

3.2.2 重复记录检测

排除了不相关记录之后，还需要检查是否有重复的文献记录。重复记录有两种情况：

1.来自同一数据库的重复记录，产生的原因是名字扩展查询导致查询结果集存在交叉。

2.来自不同数据库的重复记录，产生的原因有：a).数据库收录范围重叠 b).文献分别来自原文数据库和引文数据库。

针对这种情况，我们设计了重复记录的检测集去除和合并算法，如图 5 所示。

4 系统实现与结果分析

4.1 系统结构

为了验证上述方法的有效性，我们设计了一个系统 BibCollector。图 6 显示了系统的结构。整个系统由四部分组成，分别是：

1 抽取：系统从计算机领域常用的数据库抽取专家的文献信息。用户提交专家的名字和机构/地址信息，系统把用户请求转变成符合各数据库语法的检索表达式，向各数据库发起查询。系统只收集文献的题录信息，而并不收集文献正文。查询通过 Apache Jakarta Common 下的子项目 HttpClient5 组件和 SourceForge 的开源项目 HtmlParser6 来实现。

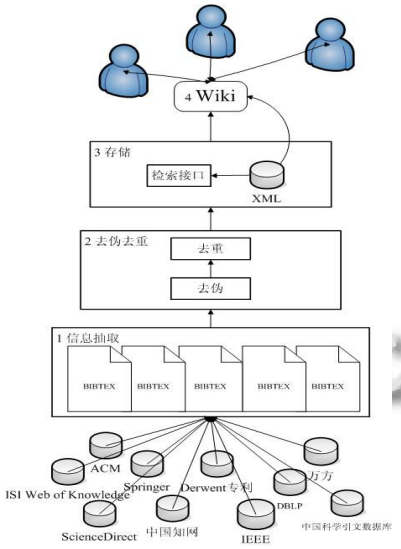


图 6 BibCollector 系统结构图

2 整合：在检索阶段完成之后，为了提高专家文献记录集的准确率，需要对文献记录进行清洗以提高数据质量。这个环节包括两个任务：过滤不相关记录，合并重复记录。不相关记录的过滤通过对比文献检索记录中的“作者”“机构”字段和查询表达式中的相应字段来实现。去重分为在单个数据库文献记录中进行的纵向去重和在来自多个数据库的记录中进行的横向去重。纵向去重使用各数据库对文献的唯一标识符来实现。我们通过 DOI(Digital Object Unique Identifier)标志来判断不同的题录是否为同一文献。DOI 是美国出版协会建立的用于唯一标识符来标识其出版的电子文献的标准。如果数据库不提供 DOI 标识，则使用该数据库自定义的唯一标识进行检测。横向去重通过综合判断两个记录的标题，出处，页码来实现。

3 存储与访问：系统收集 BibTex 格式的文献记录[13]。为了储存检索得到的题录信息，系统采用 XML(Extensible Markup Language) 可扩展标记语言进行存储[14]。XML 格式的优点是标准化，平台独立。有些数据库不提供 BibTex 格式的记录，而是提供自定义的文本信息。我们设计了相应的转换程序将文本格式的记录转变成 XML 格式。

4 信息发布：系统采取 Wiki 作为前端信息发布平台。Wiki 是多人协作平台，用户可以浏览页面，也可以参与页面的编辑维护。如果经过上述处理后的记录信息，仍然存在不准确之处时，专家作为页面的浏览者，可以直接对页面进行编辑修改。系统将更新为准确信息。

4.2 结果分析

图 7 对比了使用 ACM 主席 Stuart I Feldman 作为检索对象，从 BibCollector、DBLP、Author Profile Page、ArnetMiner 所获得的文献记录数目。可以看出，BibCollector 对 Stuart I Feldman 获取的文献记录与 ACM 数据库提供的数据量相等,说明 BibCollector 能全面准确的收集特定专家的文献记录。图 8 为使用 BibCollector 检获的 Stuart I Feldman 的文献的来源分布。

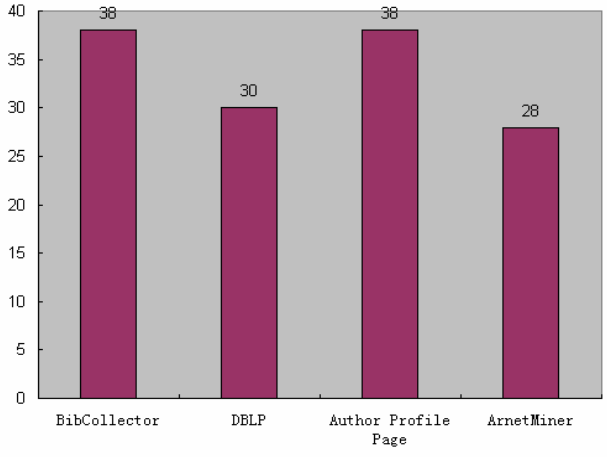


图 7 各系统检索结果对比(以 Sttuart I Feldman 为例)

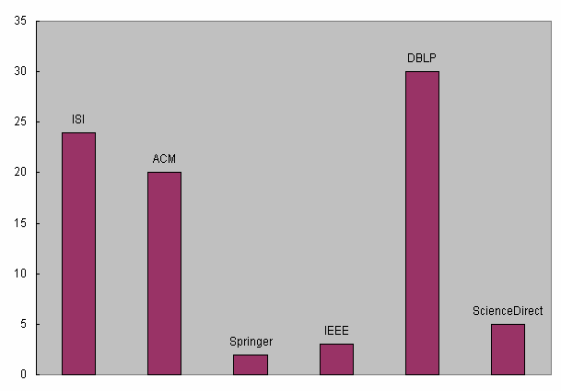


图 8 BibCollector 数据来源分布

5 总结及讨论

针对专家文献记录收集过程中所存在的问题，我们设计了一种自动文献记录收集系统。只需要提交专家的姓名和所在机构，就能够全面准确的收集文献记录。与现有的文献记录收集系统相比，我们的优点在于数据来源广泛，涵盖计算机领域常用中外全文引文数据库，并且数据准确。因为数据来源中包含了中文数据库，所以通过我们的系统可以收集中国学者的信息。本系统使用 Wiki 作为发布系统，加强用户的参与度。本系统以计算机领域为例，但可以很容易的推广到其他领域，也可以很容易推广到其他机构使用。当需要扩大数据涵盖范围时，只需要为新添加的数据库实现数据抓取程序，就能轻易实现扩展。

由于各数据库在处理地址的方法上存在差异，导致有些数据难以被检索。比如 IEEE 数据库在显示检索结果时只显示第一作者的地址，其他作者的地址需要进入全文才能浏览获得。ACM 数据库在显示检索结果时，并不显示完整的详细地址，在检索结果中只能看到国家，城市，单位，而更具体的部门一级的信息则要在全文中才能看到。这些情况的存在，都为获得精确和全面的检索设置了障碍，需要进一步的考虑，同时也凸显了制定各数据库统一的命名及地址规范的需要。

参考文献

1 Amit Singhal. Modern Information Retrieval: A Brief Overview. IEEE Data Engineering Bulletin. New York; IEEE. 2001:35–43.

2 Ricardo Baeza-Yates, Berthier Ribeiro-Neto. Modern Information Retrieval. New York: ACM Press/Addison Wesley Longman, 1999.

3 周津慧,王衍喜,王永吉,关贝,郝丹.基于领域专家学科知识链的文献资源组织与导航.科研信息化技术与应用,2011,2(1):33–42.

4 王衍喜,周津慧,王永吉,肖永红,郝丹.一种基于科技文献的学科团队识别方法研究.图书情报工作,2011,55(2):55–58.

5 Michael Ley. The DBLP computer science bibliography: evolution, research issues, perspectives. String Processing and Information Retrieval, 9th International Symposium, SPIRE. Lisbon, Portugal; Springer, 2002:1–10.

6 Michael Ley. DBLP-Some lessons learned. Very Large Data Base. VLDB Endowment. 2009:1493–1500.

7 Michael Ley, Patrick Reuther. Maintaining an online bibliographical database: the problem of data quality. Proc. of the Extraction et Gestion des Connaissances. Lille, France, 2006. Cepadues-Editions.5–10.

8 Tang Jie, Zhang Jing, Zhang Duo, Yao Limin, Zhu Chunlin, Li Juanzi. ArnetMiner: An expertise oriented search system for web community. Proc. of the 6th International Conference of Semantic Web. Graz, Austria, 2007. ACM New York,1–8.

9 Hui Han, Hongyuan Zha. Name disambiguation in author citations using a K-way spectral clustering method. Proc. of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries. Denver, CO, USA; ACM. 2005:334–343.

10 Han H, Giles L, Zha H, Li C, Tsioutsiouliklis K. Two supervised learning approaches for name disambiguation in author citations. Proc. of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries. Tucson, AZ, USA, 2004. 296–305.

11 Jeremy A. Hylton. Identifying and merging related bibliographic records[Thesis (M. Eng.)]. Massachusetts; Massachusetts Institute of Technology, 1996.

12 郝丹,周津慧,关贝,王衍喜,韩继欣.文献跨库检索中去重方法研究与应用.现代图书情报技术,2011,7(8):116–120.

13 J Fenn. Managing citations and your bibliography with BibTeX. The PracTeX Journal, 2007,4(1):1–19.

14 Luca Previtali, Brenno Lurati, Erik Wilde. BIBTEXML: An XML representation of BibTex. Proc. of the Tenth International World Wide Web Conference. Hongkong, China, 2001. ACM Press.64–65.