基于推理机的 SCI 地址字段数据清洗方法设计

张晋辉[1,2],刘 清[2]

(1.中国科学院 研究生院,北京 100190;
2.中国科学院 国家科学图书馆武汉分馆,湖北 武汉 430071)

摘 要:探讨了将推理机引入到 SCI 地址字段数据清洗中的方法。首先通过指出目前 SCI 地址字段数据清洗方法的不足阐述了进行 SCI 地址字段数据清洗方法研究的必要性, 然后介绍了推理机的基本原理,并对应用于 SCI 地址字段数据清洗中的推理机进行了设计,包括待推理数据的生成、知识库的构建及推理控制策略的设计等,旨在提出适用于 SCI 地址字段数据清洗的方案。

关键词:推理机;SCI 地址字段;数据清洗;方法;设计

中图分类号:G350 **文献标识码**:A **文章编号**:1007-7634(2010)05-0741-06

Design of A Data Cleaning Method of SCI Author Addresses Based on Inference Engine

ZHANG Jin-hui[1,2], *LIU Qing*[2]

(1.*Graduate University of Chinese Academy of Sciences, Beijing* 100190, *China;*
2.*Wuhan Branch of the National Science Library, CAS, Wuhan* 430071, *China*)

Abstract: Aiming on improving the effectiveness and reliability of specific bibliometrics application, this paper introduces inference engine into the data cleaning procedure which is suitable for SCI author's addresses field. By illustrating the limitations of the current data cleaning methods with address field, this paper has accentuated the necessity of a better way in cleaning. After intrducing the basic principles of inference engine, the paper takes the articles of Chinese Academy of Sciences in SCI database as an example to design an inference engine used in the data cleaning of SCI author's addresses, including the reasoning data table, the knowledge database and the reasoning control strategy and so on. The advantages and disadvantages of this method are also summerized.

Keywords: inference engine; SCI Institutional addresses; data cleaning; method; design

1 引 言

数据清洗源于数据仓库、数据挖掘等领域,指的是从数据源中检测和消除错误数据、不一致数据和重复数据,从而改善数据库中数据质量的过程。虽然数据清洗得到了国内外学者的普遍关注, 然而截止到现在,它也没有公认的定义,不同的应用领域对其有不同的解释[1-4],但其本质是不变的,就是消除"脏数据"的过程。这里的"脏数据"是不同应用领域数据源中不符合要求的数据的统称。后来,数据清洗被广泛应用于和数据处理相关的各个领域, 也包括文献计量领域。

文献计量工作中,数据清洗是一个非常重要的、不可或缺的环节,应用了众多的方法和手段。这些方法和手段以人工和半自动为主。其中,人工清洗结果精确得当,但工作量浩繁无比,效率甚低;而半自动清洗在速度上提高不少,但清洗结果可靠性差,错误

收稿日期:2009-04-20

作者简介:张晋辉(1981-),男,河北秦皇岛人,硕士研究生,主要从事学科情报研究;刘 清(1969-),男,湖北武汉人,研究员,硕士生导师,主要从事情报学理论与方法研究.

繁多(关于这一点有很多原因,后文将有阐述),需人工辅助再次清洗。按照我们的实际工作经验,文献计量分析工作中,数据清洗所占的时间占全部工作量的 80–90%。由此可以想见好的、高效的数据清洗方法对于文献计量工作的重要性是如何之大。

推理机(Inference Engine)是一种通用性的知识处理系统,其基本原理是在一定的控制策略指导下搜索规则库中的可用规则,和动态数据库匹配从而得出推理结论。文献计量领域中的数据清洗是情报分析人员在数据分析之前对原始数据进行整理或预处理的过程。该过程实际上就是一个推理过程,与推理机工作原理相似。它们都需要基于领域专家的知识对原始数据进行推理,这为我们将推理机应用于文献计量工作中的数据清洗提供了可行性。因此,本文试图将推理机方法引入文献计量工作中的数据清洗环节,一方面提升数据清洗的速度,另一方面提高数据清洗结果的精确度和可靠性。

目前文献计量领域最为常用的分析数据源《科学引文索引》(Science Citation Index,以下简称 SCI),包含的众多字段为不同类型的分析需求提供了大量高价值的信息,其中地址字段(Addresses)在文献计量应用中是非常重要的,被频繁分析的对象之一。所以本文选取 SCI 中文章的地址字段作为提出问题和解决问题的对象,设计将推理机应用于 SCI 地址字段数据清洗中,以探索数据清洗的新方法,为其他数据源和字段提供基于推理机的清洗提供参考。

2 目前 SCI 地址字段数据清洗方法的不足

情报研究人员主要采用人工方式和半自动方式进行数据清洗。人工方式存在耗费大量人力、财力和物力以及效率低下的弊端,笔者不再详细阐述。半自动方式主要指的采用那些提供清洗功能的数据分析软件(如 TDA)进行数据清洗的方式。

TDA 即 Thomson Data Analyzer,是汤姆森公司推出的科技信息分析工具,也是目前普遍使用的一个数据分析软件,同时具有强大的数据清洗功能,为情报分析人员进行数据整理提供了一个非常方便的工具。在使用 TDA 进行数据清洗时我们主要使用它的 list cleanup 和 thesauri 工具,list cleanup 基于模糊匹配算法,调用相关模块对相应字段进行自动整理;thesauri 是根据实现设定好的叙词表对字段进行整理,叙词表根据具体处理字段分为多种(如国家词表、作者词表等),叙词表中存储了相近词的聚类,它将某词的相近词归并到该词下面,一旦记录中出现该词或该词的相近词,便统一划归到该词下面,从而实现清理过程。

TDA 较之手工方式进步了很多,但这主要体现在效率方面,而在其它方面有很多不足之处。它存在很多分析和处理错误,如笔者在 SCI 的标题字段选取"water pollution",地址字段选取"Peoples R China"进行检索,得到 104 条数据结果,再利用 list cleanup 工具对其地址字段进行清洗得到结果,如图 1 所示。

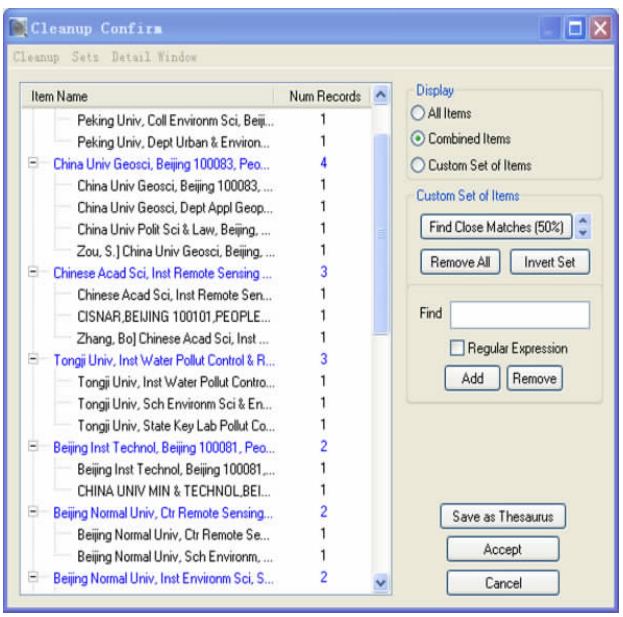

图 1 TDA list cleanup 数据清洗结果展示图

图 1 中可以发现它将中国政法大学归并到了中国地质大学下面,又将 CISNAR(北京大学东北亚区域一体化研究中心)合并到了中科院遥感应用研究中心的下面,从而产生了很明显的分析错误,而这仅仅是一次检索中从少量数据中发现的问题,对于情报研究工作中的大规模数据集,TDA 的局限性可想而知。究其原因是,list cleanup 中的判定模块主要基于模糊匹配算法识别词形相近的词,而对概念相近的词却无能为力,它并不能对两个字段的概念关系进行判断,只是停留在词形上的简单比对上。虽然可以解决一些因误拼或变形等造成的不一致字段的数据整理问题,但不能根本上解决复杂的文献计量数据清洗问题,而叙词表的设计比较简单,仅仅将与某词或词组形式上相关的单元归并到该词或词组下面,不能对概念相关词进行准确判断。也正是考虑到这一点,TDA 提供了 Cleanup Confirm 界面,允许用

户进行手工整理，但这又需要耗费大量的精力和时间去修正 TDA 的分析错误。同时叙词表的数量不能满足丰富的分析需求，比如它没有提供机构叙词表，而如果都要用户进行设计对用户的专业知识和时间、精力等是个极大的挑战。从根本上解决概念相似词判断的问题，必须借助领域专家的知识、基于专家知识的判断设计更为高效、快捷的数据清洗方法。下面我们就来介绍一下基于专家知识的推理机的原理并探讨将其应用于 SCI 地址字段数据清洗的可能性。

3　推理机及其基本原理

推理机(Inference Engine) 是专家系统的思维机构，是构成专家系统的核心部分，能够模拟领域专家的思维过程，从而控制并执行对问题的求解。它是一种通用性的知识处理系统，其基本原理是在一定的控制策略指导下搜索规则库中的可用规则，和动态数据库匹配从而得出推理结论。推理机的工作原理如图 2 所示，其中动态数据库存储用户对领域知识或客观事实的描述词汇，即供推理的基本事实、推理的中间结果、最终结论等有关信息的，供推理机用于和规则知识库中的规则条件进行匹配，以得出推理结论；规则库中存储的是一种经过分类组织的知识集合，包含领域专家的事实知识、经验知识等，即用于存储推理使用的基本规则，需要相关领域专家根据实践经验总结形成；推理控制机制是推理机的大脑和核心，负责将动态数据库中的原始事实与规则库中的推理规则进行匹配并得出结论。

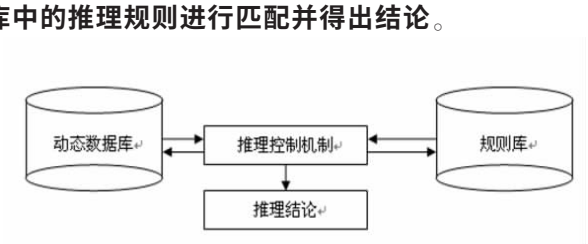

图 2 推理机工作原理示意图

从图 2 中我们可以看到，推理机运行的关键是规则库的构建和推理控制策略的制定，规则库构建的合理与否、推理控制策略制定的完备与否直接决定了推理机得出结论的效率和准确性。规则库的构建基于领域专家的知识，根据专家实践经验的长期积累得出，而推理控制机制的制定则需要考虑推理方式确定、推理的实现过程设计等诸多因素。其中推理方式根据推理方向的不同可分为正向推理、反向推理及双向推理三种[5]。正向推理是从已知事实出发，正向使用推理规则，就是使用动态数据库中的事实去匹配规则库中的规则的前提条件，从而得出推理结论，是一种数据驱动的推理方式；反向推理是一种以某个假设目标为出发点，反向运用推理规则的推理方式，是一种目标驱动的推理方式；双向推理即正、反向推理的混合使用。

文献计量领域中的数据清洗是情报分析人员在数据分析之前对原始数据进行整理或预处理的过程，即分析人员首先从文献计量数据源中检索得到原始数据，然后根据文献计量专家的实践经验和已有知识对原始数据进行判断并整理，得到符合要求的数据即便于统计、分析的数据，该过程如图 3 所示。

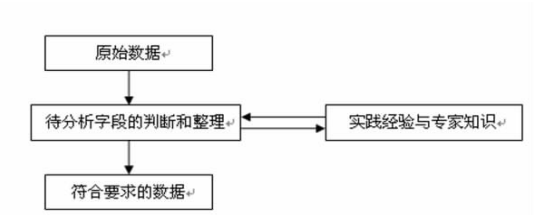

图 3 文献计量领域数据清洗过程示意图

从图 3 中我们可以发现，该过程实际上就是一个推理过程，与图 2 的推理机工作原理图存在很大的相似性，它们都需要基于领域专家的知识对原始数据进行推理，这为我们将推理机应用于 SCI 地址字段的数据清洗提供了可行性，并且推理机在其它领域得到了成功的应用，所以推理机也完全可以应用于文献计量数据清洗领域中。

4　SCI 地址字段数据清洗推理机设计

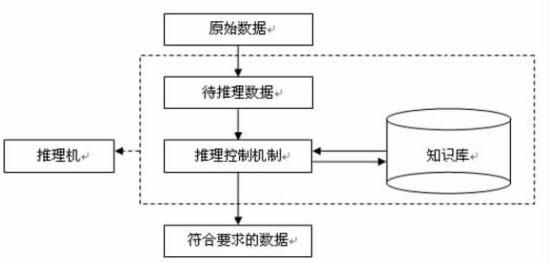

图 4 推理机应用于 SCI 地址字段数据清洗的总体框架图

根据推理机的工作原理，我们将推理机应用于 SCI 地址字段数据清洗的总体框架描述如下：首先从文献计量数据库中检索取得原始数据，并对其加工处理形成待推理数据，然后推理机在知识库中相关规则的控制下对待推理字段进行推理判断，并将结论存入结果表中，完成清洗，得到符合要求的数据，如图 4 所示。SCI 地址字段数据清洗推理机工作的前提是待推理数据的生成、推理知识库的构建以

及推理控制策略的设计，下面我们就对这些问题展开进行讨论。

4.1 SCI 地址字段数据清洗推理机中待推理数据的生成

SCI 地址字段数据清洗推理机中的待推理数据需要结合用户的具体分析需求，对原始数据进行加工处理，抽出待推理字段（本文中为地址字段）存储起来，其基本结构可以用表格的形式展现。其中字段序号对应原始记录的记录号，待推理字段需要对原始记录进行处理后得出。我们以中科院文章地址字段的待推理数据表为例进行说明，如表 1 所示。

表 1 待推理数据表

字段序号	待推理字段
1	Chinese Acad Sci, Inst Semicond, Lab Nanooptoelect, Beijing 100083, Peoples R China
2	Chinese Acad Sci, Inst Semicond, SKLSM, Beijing 100083, Peoples R China
3	CAS, Inst Geog Sci & Nat Resources Res, Beijing 100101, Peoples R China
…	…
I	…

4.2 SCI 地址字段数据清洗推理机知识库的构建

知识库中的知识即规则必须有适当的表示才便于存储、检索、使用和修改，知识的表示就是指如何用最合适的形式，对问题进行中所需要的各种知识进行组织。知识的表示方式有很多种[6]，笔者建议采用产生式规则[7]表达方式对文献计量知识库中的知识进行表达。产生式知识表示方式是专家系统中最常见也是最容易理解的知识表示方式。它采用的“如果…则…”知识表示形式与人们的思维习惯很类似，同时又容易在计算机中实现。这种知识的表示方法具有自然、直观、便于实现推理的优点[8]。产生式规则一般形式如下：

IF Px

THEN Cn

其中 Px=P1 AND/OR P2…Pn。P 表示条件，它既可以是一个简单条件，也可以是由多个简单条件构成的逻辑组合，例如 P=P1 AND P2 OR P3。Cn 表示结论，它可以是一个或多个结论，P 和 C 的值需要具体领域的专家根据经验并经长期验证给出。

SCI 地址字段数据清洗推理机知识库中的规则同样需要文献计量领域的专家结合长期实践经验不断总结得出，并且该知识库是随着推理机的运行而不断丰富的，即起初推理机的知识库不可能包括所有规则，在一次推理完成后待推理数据中的一些记录没有得到推理结论，这就需要文献计量专家的干预，从没有得到推理结论的记录中再次提炼出规则存储到知识库中，作为下次推理使用的规则，这样周而复始不断重复该过程，知识库得到不断丰富。

我们以中科院文章地址字段的判定为例对如何利用产生式规则描述 SCI 地址字段的知识进行说明，通过对大数据集的分析发现地址字段中有些条件是共性的，这里的条件指的是地址字段中逗号分隔出的单元，比如第一条记录：“Chinese Acad Sci, Inst Semicond, Lab Nanooptoelect, Beijing 100083, Peoples R China”可分隔为“Chinese Acad Sci”、“Inst Semicond”、“Lab Nanooptoelect”、“Beijing 100083”、“Peoples R China”等五个单元。当字段中出现某个单元或单元的组合时便可以判断出文章机构来源。如当地址字段中存在 CAS 和 Peoples R China，或者存在 Acad Sinica 和 Peoples R China，或者存在 Chinese Acad Sci 或 China Acad Sci 时就可以判断该记录是中国科学院的文章，下面我们用产生式规则进行描述。

规则 1：IF （CAS OR Acad Sinica AND Peoples R China）OR Chinese Acad Sci OR China Acad Sci THEN 中国科学院，这是一级机构的判定，下级机构的判定比较复杂，但原理相同，以中科院半导体所为例，描述如下：

规则 2：IF ((CAS OR Acad Sinica AND Peoples R China)OR Chinese Acad Sci OR Acad Sinica OR China Acad Sci)AND （Inst Semicond OR Lab Nanooptoelect OR Key Lab Semicond Mat Sci OR Novel Mat Lab OR Nanooptoelect Lab OR Key Lab Semicond Mat OR NLSM OR Nano Optoelect Lab OR Novel Semicond Mat Lab OR State Key Lab Superlattices & Microstruct OR SKLSM OR State Key Lab Integrated Optoelect OR State Key Lab Superlattice & Microstruct OR Natl Lab Superlattices & Microstruct OR Semicond Lighting Technol Res & Dev Ctr OR Lab Semicond Mat Sci OR Res & Dev Ctr Semicond Lighting Technol OR Natl Key Lab Integrated Optoelect OR State Key Lab Superlatt & Microstruct OR Natl Res Ctr Optoelect Technol)

THEN 中科院半导体所

具体到这些规则在知识库中的存储结构，我们可以设计几张表，用表格的形式进行存储，我们将条件表分为国家表、一级机构表、二级机构表和三级机

构表等，分别用于存放国家、一级机构、二级机构和三级机构的对应条件。规则表中存储着条件组合和结论号，结论号与结论表中的结论号对应，根据结论号可以从结论表中识别出结论内容，条件组合中的值由条件表中的条件编码组合而成，它组成了规则中的条件部分。条件表中的条件编码的形式根据条件表的不同而不同；国家表中我们采用两个大写字母代表国家，如 CN 代表中国，US 代表美国等；一级机构表中用大写字母 Z 代表中国科学院，后面跟两位数字以唯一标识一级机构判定规则条件部分中的某个单元；同理二级或三级机构表中在大写字母 Z 后面跟三位或四位数字以唯一标识二级或三级机构判定规则条件部分中的某个单元；结论表用于存放中科院的所有机构及其对应的结论号，我们将其设计如表 2–表 7 所示。

表 2 国家表

条件号	条件内容	条件编码
1	Peoples R China	CN
2	USA	US
…	…	…
n	…	…

表 3 一级机构表

条件号	条件内容	条件编码
1	Chinese Acad Sci	Z01
2	CAS	Z02
3	Acad Sinica	Z03
4	China Acad Sci	Z04
…	…	…
n	…	…

表 4 二级机构表

条件号	条件内容	条件编码
1	Inst Semicond	Z001
2	Beijing Inst Genom	Z002
3	Inst Geog Sci & Nat Resources Res	Z003
…	…	…
n	…	…

表 5 三级机构表

条件号	条件内容	条件编码
1	Lab Nanooptoelect	Z0001
2	Key Lab Semicond Mat Sci	Z0002
3	Novel Mat Lab	Z0003
…	…	…
n	…	…

表 6 规则表

规则序号	条件组合	所属地点	结论号
1	CNZ02	Beijing	1
2	CNZ03	Beijing	1
3	Z01	Beijing	1
4	Z04	Beijing	1
6	CNZ02Z001	Beijing	2
7	CNZ02Z002	Beijing	2
8	CNZ02Z003	Beijing	2
…	…	…	…
J	…	…	…

表 7 结论表

结论号	结论内容
1	中国科学院
2	中国科学院半导体所
3	中国科学院北京基因组研究所
…	…
n	…

4.3 SCI 地址字段数据清洗推理机的推理控制策略

SCI 地址字段的数据清洗中，原始数据需要经过推理机的判断得到推理结论，事先我们并不知道结论，所以这里推理机采用的主要是正向推理方式，其基本思想是，用户事先从文献计量数据源中检索得到原始数据并将其中待推理的字段抽出存储到待推理数据表中，然后由推理机工作得出推理结论存储到结果表中，从而完成对原始数据的推理和整理。

我们将 SCI 地址字段数据清洗推理机的推理过程设计为：推理机逐一取出待推理数据表中的待推理字段，并将待推理字段按逗号分隔成单元存储到词组中，逐条扫描知识库中的规则，从知识库中取出规则中的条件组合，再从条件表中找出条件组合中对应的条件内容，然后判断该条件组合中的条件内容是否在待推理字段形成的词组中，如在就取该规则的结论号，到结论表中找出其对应的结论内容并赋值到结果表中的结论字段中，如不在就取出下一条规则的条件组合进行判断，直至目标出现或知识库中再无可用的规则为止。一条记录处理完就转到待推理数据表中的下一条记录，如此循环执行，直至完成对整个待推理数据表的扫描。我们用流程图描述整个推理过程，其中 i 为待推理数据表中的字段序号，j 为知识库中的规则序号，I 为待推理数据表中的记录总数，J 为规则库中的规则总数（图 5）。

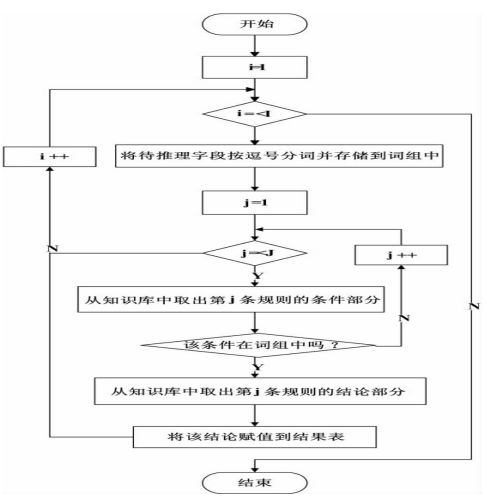

图 5 推理机工作流程图

根据此推理过程，我们仍以中科院地址字段清洗进行说明，推理机首先从待推理数据表（本文为表2）中取出第一条记录中的待推理字段“Chinese Acad Sci, Inst Semicond, Lab Nanooptoelect, Beijing 100083, Peoples R China”，并将此字段按逗号分隔成单元存储到词组1中，然后判断其字段序号是否小于待推理数据表中的总记录数I，如果不是则直接退出，如果是则取出规则表中的第一条规则中的条件组合“CNZ02”，并判断规则序号与规则总数的关系，如果大于J则说明规则库中所有规则都不适应于该记录，则取出其下一条记录重新进行判断，如果不大于J则找出其对应的条件内容“CAS”和“Peoples R China”，判断其是否在词组1中，显然不在，然后取出下一条规则直至规则中的条件组合为“Z01Z001”或者“Z01Z0001”或者其它的适用规则，其条件内容为“Chinese Acad Sci”和“Inst Semicond”或者“Chinese Acad Sci”和“Lab Nanooptoelect”或其它，判断它们存在于记录中，然后取出规则表中的记录号“2”，找到其对应的结论内容“中科院半导体所”赋值到该结果表中的结论字段中，这样就完成了第一条记录的推理过程，然后依此类推，直至完成对整个待推理数据表的扫描，最终形成一张结果表，如表8所示。结果表是在原来待推理数据表的基础上添加结论字段形成的，结论字段由推理机完成推理后将结论赋值过来。

表8 结果表

字段序号	待推理字段	结论
1	Chinese Acad Sci, Inst Semicond, Lab Nanooptoelect, Beijing 100083, Peoples R China	中科院半导体所
2	Chinese Acad Sci, Inst Semicond, SKLSM, Beijing 100083, Peoples R China	中科院半导体所
3	CAS, Inst Geog Sci & Nat Resources Res, Beijing 100101, Peoples R China	中科院地理科学和资源研究所
…	…	…
n	…	…

5 结 语

本文针对SCI数据库中的作者地址字段进行了基于推理机的清洗方案设计，具体实施还有待下一步的编程实现。对于一些技术细节比如如何将待推理字段分词、如何将规则表中的条件与待推理字段对应的词组进行匹配、匹配后如何赋值等等都尚待进一步研究。

当然，这一方案也存在一些问题。比如随着清洗操作的不断进行和规则的不断增加，知识库会逐渐增大，从而影响到清洗效率。实际应用中，还应结合计算机编程人员，对知识库和清洗程序进行科学地设计，从技术上提高清洗效率。

此外，本文提出的数据清洗实际上还包括数据的整理。严格来说，数据清洗只是针对“脏数据”，是不针对规范和清洁的数据的。但是对于文献计量的实际应用来说，为便于统计和计量的目的，需要对所有数据进行整理，所以这里的数据清洗包含数据整理。这在一定程度上也影响到清洗的效率。但是，考虑到实际应用的需求，这种方案对于整体效率的提高实际上还是有很大帮助的。

第三，本文仅仅针对SCI数据库中的作者地址字段进行了方案设计。由于其它数据库和其它字段与SCI的地址字段存在差别，直接照搬并不太可行，但是由于基本原理基本相通，基于推理机的数据清洗方法在文献计量领域的应用还是存在实用价值的。

参考文献

1 梁文斌.数据仓库中数据清洗的研究与设计[D].苏州：苏州大学，2005.

2 苏 成.数据挖掘中不可忽视的环节——数据预处理[J].华南金融电脑，2006，(1)：64-66.

3 Simoudis E，Livezey B and Kerber R.Using Recon for Data Cleaning[A].Proceedings of KDD[C].1995：282-287.

4 Levitin A and Redman T.A Model of the data (1ife) cycles with application to quality[J].Information and Software Technology，1995，35(4)：217-223.

5 崔 萌. 专家系统推理机核心设计[J].中国高新技术企业，2008，(22)：298-299.

6 丁全龙，吴保国.一种基于产生式规则的造林专家系统的设计与实现[J].农业网络信息，2006，(8)：16-18.

7 王秀和，李国建，赵振卫.基于一种新知识库结构的电机设计专家系统推理机[J].微特电机，2000，(4)：8-14.

8 尹朝庆，尹 皓.人工智能与专家系统[M].北京：中国水利水电出版社，2001：37-61.

（责任编辑：徐 波）