# 基于非相关文献的三阶知识发现方法探讨*

李 勇 冷伏海 王 林

**摘 要** 在目前非相关文献知识发现的开放式发现模式的基础上作进一步的共现发现,即通过三个文献集的分析能够推导出一种潜在的新关系,这种新关系是通过单独分析一个或两个文献集所无法获得的。通过在激光显示领域的实证研究,发现了一条"从实际需求出发—实现这一应用的技术系统—提高这一系统性能的方法—为该方法提供理论支撑的理论"的潜在关联链条,验证了基于非相关文献的三阶知识发现方法可以以研究目标和具体问题为导向,为研究活动提供服务。图4。表4。参考文献9。

**关键词** 知识发现 非相关文献 三阶共现 关联传递 知识服务

**分类号** G350

**ABSTRACT** On the basis of the current theory and practice of disjoint literature-based discovery, this paper explores the third-order disjoint literature-based discovery method. The framework and process of the third-order disjoint literature-based discovery method are proposed. Furthermore, a preliminary experiment is carried out in the field of laser display and a potential chain from the actual needs and technology system to theory is found. The third-order disjoint literature-based discovery method can be goal and problem-oriented and thus be served in supporting research activities. 4 figs. 4 tabs. 9 refs.

**KEY WORDS** Knowledge discovery. Disjoint literature. Third-order co-occurrence. Relationship relay. Knowledge services.

**CLASS NUMBER** G350

## 1 研究背景

1986年,美国芝加哥大学的Don R. Swanson教授首次提出"基于非相关文献的知识发现法"[1-2]。所谓基于非相关文献的知识发现法就是从表面上没有任何联系的文献内容之间识别出有效的、新颖的、潜在有用的以及最终可理解的知识的情报研究方法。该方法可辅助科研人员发现潜在的关联,进而促进新知识的产生,有助于推动科学的发展[3]。

该方法经过多年的发展,在很多方面取得了很大进步[4]。但是目前的研究均基于Swanson提出的非相关文献知识发现的基本发现模式,即以感兴趣的主题A为初始点,生成集合A,将主题A共同出现在题名中的术语列举出来,排除停用词汇、通用词汇后形成一个新的"潜在发现"的词表B,并对词表按照某一阈值进行排序。将与B集合中的词汇同时出现在题名中的词汇汇集起来,形成集合C。通过发现集合B与集合C之间的潜在联系,最终发现A与C之间的关系。

而如果我们在非相关文献知识发现的初始文献集(文献集A)构建、信息抽取和对中间集(文献集B)过滤与排序后,对得到的目标集(文献集C)进行过滤和排序,并以文献集C的主题作进一步的共现发现,会得到什么样的结果?即现有非相关文献知识发现是经过二次关联关系传递的二阶知识发现,经过三次及多次关联关系传递的是三阶、多阶知识发现,这种发现是否有意义,操作可行性如何?通过文献调研,在已经公开发表的文献中,未发现有对基于非相关文献的多阶知识发现方法的原理、模式及流

* 本文系国家自然科学基金项目"科技创新演化分析理论与方法研究"(编号:70873123)的研究成果之一。

程进行研究和实证的论文。本文将首次对基于非相关文献的三阶知识发现方法进行探讨。

## 2 相关研究

在社会学的一项研究中[5-6]，美国弗吉尼亚大学的计算机专家 Brett Tjaden 设计了一个程序“Game of Kevin Bacon”，以电影演员 Kevin Bacon 为中心，定义了“Bacon Number：对其他演员”，如果他（她）和 Kevin Bacon 一起演过电影，则其 Bacon Number 为 1；如果他（她）没有和 Bacon 演过电影，但是和 Bacon Number 为 1 的演员一起演过电影，则其 Bacon Number 为 2，以此类推。即通过是否共同出演一部电影来构建演员间的共现关系。研究人员统计了 585,220 个演员及 275,000 部电影的信息。表 1 是对所有演员做的统计。左边是 Bacon Number，右边是 Bacon Number 为该值的演员数。平均 Bacon Number 为 2.944。

**表 1 Bacon Number 的多阶共现统计**

| Bacon 数 | 演员数 | 累积百分比 |
|---|---|---|
| 0 | 1 | |
| 1 | 1,682 | 0.287 |
| 2 | 132,399 | 22.91 |
| 3 | 357,230 | 83.95 |
| 4 | 86,206 | 98.68 |
| 5 | 6,734 | 99.83 |
| 6 | 852 | 99.98 |
| 7 | 103 | 99.99 |
| 8 | 13 | 100.00 |

如表 1 所示，在此项研究中一阶共现的仅占 0.287%，二阶共现的占 22.91%，三阶共现的占 83.95%，四阶共现的为 98.68%。

若将 Bacon Number 中的演员理解为主题，并将电影理解为文献，则可将此研究视作一次基于非相关文献的多阶知识发现，即一阶共现（直接共现）的仅占 0.287%，二阶共现（现有的非相关文献知识发现方法）的占 22.91%，三阶共现的占 83.95%（基于非相关文献的三阶知识发现方法），四阶共现的为 98.68%，五阶共现的为 99.83%。也就是说在此次知识发现研究中，二阶共现的发现仅占全部非直接共现的 22.9%，三阶或四阶的发现仍有很大的意义。

以上研究表明，基于非相关文献的多阶（三阶、四阶……）知识发现方法值得探讨，本文首先提出了基于非相关文献的三阶知识发现的基本原理、发现模式，并尝试在激光显示领域进行实证。

## 3 基于非相关文献的三阶知识发现方法的基本原理

基于非相关文献的三阶知识发现方法的基本原理是：如果一组文献的集合描述了主题 A 和 B 的关系，一组文献的集合描述了主题 B 和 C 之间的关系，而尚没有关于主题 A 和 C 关系的报道，则两个文献集之间可能存在潜在的关联；另一组文献的集合描述了主题 C 和 D 之间的关系，同样尚没有关于主题 A 和 D 或主题 B 和 D 关系的报道，则文献集 A 和 D 之间也可能存在潜在的关联，即通过对这样三个文献集的分析能够推导出一种潜在的新关系，而这种新关系是通过单独分析一个或两个文献集所无法获得的。

由于主题 A 和 C 或主题 B 和 D 的关系均没有报道，因此在上述过程中，通过分析文献集 A 和 B 或文献集 B 和 C 也同样可以分别推导出潜在的新关系，这些新关系也是通过单独分析一个文献集所无法获得的。

基于非相关文献的三阶知识发现方法的基本框架可概括为图 1。发现过程始于一个研究人员感兴趣的初始主题 A，然后构建初始“文献集 A”；通过对关键词进行提取、排序、过滤，得到表征 A 概念的有序关键词列表，其中的每一个关键词称为“B 概念”，所形成的文献集称为“文献集 B”；然后对 B 文献集重复上述数据处理形成有序词表，并与“B 概念”对比去重，得到“C 概念”的有序词表并形成“文献集 C”；再次处理后形成“目标概念”或“D 概念”的有序关键

词列表,这样就为初始词提供了一个有序的可能存在潜在关联的词汇列表。

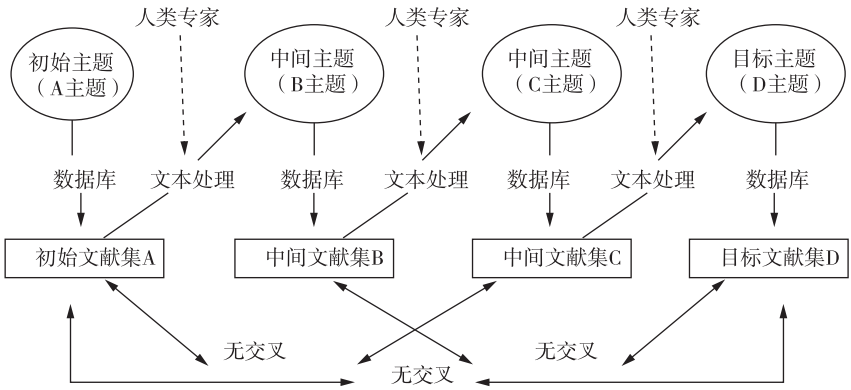


**图1 基于非相关文献的三阶知识发现方法的基本框架**

从以上分析看出,基于非相关文献的三阶知识发现方法可以发现更多的潜在关联,这意味着有更多的潜在知识等待去挖掘。同时三阶的知识发现相对于二阶的知识发现有更为庞大的中间词表、更多的中间关联(三次关联 vs · 二次关联),这对数据集构建、信息抽取、中间集的排序过滤算法等技术提出了更高的要求。

## 4 基于非相关文献的三阶知识发现模式

在基于非相关文献的三阶知识发现模式中,考察其关联关系传递的方向和路径,如图2所示。

图2中,以实线表示直接关联关系,虚线表示潜在关联关系,可以发现基于非相关文献的三阶知识发现过程中存在3次直接关联关系(A→B, B→C, C→D)和3次潜在关联关系(A→C, B→D, A→D)。根据非相关文献三阶知识发现方法的基本原理,关联关系 A→B、B→C、C→D、A→C、B→D、A→D 存在多种组合,可构成非相关文献的三阶知识过程。

根据具体发现过程中的关联传递关系及产生潜在关联的不同,又可将发现模式细分如下:

(1)经过三次直接关联关系传递的三阶知识发现模式(1+1+1模式),产生一次假设:通过构建直接关联关系 A→B、B→C、C→D,最终发现潜在关联关系 A→D。

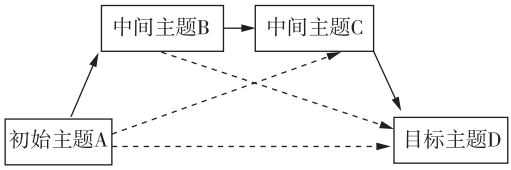


**图2 发现模式中的关联关系传递**

(2)首先经过一次潜在关联关系,然后经过一次直接关联关系传递的三阶知识发现模式(2+1模式):首先通过构建直接关联关系 A→B、B→C,发现潜在关联关系 A→C,然后构建直接关联关系 C→D,最终发现潜在关联关系 A→D 存在,即通过潜在关联 A→C 和直接关联 C→D 发现潜在关联关系 A→D。

(3)首先经过一次直接关联关系传递,然后经过一次潜在关联关系传递的三阶知识发现模式(1+2模式):通过构建直接关联关系 A→B、B→C、C→D,发现潜在关联关系 B→D,最终发现潜在关联关系 A→D,即通过直接关联 A→B 和潜在关联 B→D 发现潜在关联 A→D。

以下以"1+1+1模式"为例,分析其关联路径发现模式。

### 4.1 “1+1+1模式”的关联路径及发现模式

首先通过构建直接关联关系A→B、B→C、C→D,最终发现潜在关联关系A→D。其传递方向和路径如图3所示:

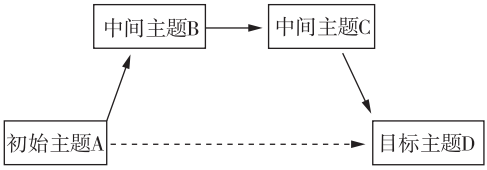


图3 “1+1+1模式”的关联关系传递

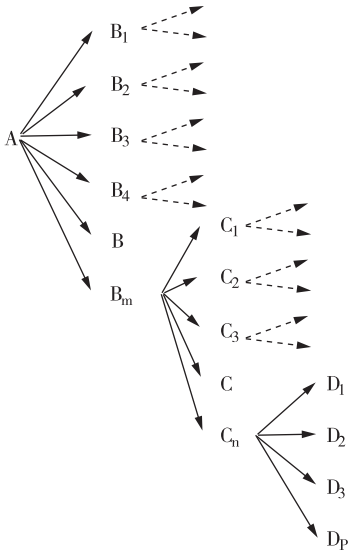


图4 基于非相关文献的“1+1+1模式”

该发现模式(见图4)始于一个研究人员感兴趣的主题,并用词或短语表示该主题的概念,即“A概念”,然后在数据库中,将所有包含A概念的文献下载,形成初始“文献集A”。利用信息抽取技术,从初始文献集中将表征A概念的词或短语抽取出来,经过过滤形成“中间概念”或“B概念”的有序词表,所形成的文献集称为“中间文献集”或“文献集B”;对B文献集重复上述数据处理,得到“C概念”并形成“文献集C”;对C文献集重复上述文本处理过程,并在人类专家的干预下,得到“目标概念”或“D概念”。

最后,一个通过B和C连接A和D的潜在关联产生,其具体的发现模式见图4。

### 4.2 发现模式分析

(1)在“1+1+1模式”中,如果关联关系A→C为直接关联关系时,即通过概念A可直接发现C,则说明构建直接关联关系A→B过程中的过滤排序算法存在不足,将本来可以通过A直接发现的部分关联过滤掉了。

同理,如果关联关系B→D为直接关联关系时,即通过概念B可直接发现D,则说明构建直接关联关系B→C过程中的过滤排序算法存在不足,将本来可以通过B直接发现的部分关联过滤掉了。

(2)反方向考虑,如果能将潜在关联的A→C和B→D转为直接关联关系,那么三阶知识发现就会转化为二阶知识发现,进而提高发现效率。

(3)“2+1模式”和“1+2模式”的中间潜在关联在理论上可以实现中间主题的进一步收敛,有助于提高整体发现效率。

## 5 应用前景

(1)首先,基于非相关文献的三阶知识发现方法仍然是一种获取同已知文献存在隐含关联的非相关文献的手段,这是与现有的基于非相关文献的二阶知识发现方法存在共性的一点。但是由于多了一次关联传递,如图4中所示,假设$B_x$($x\in[1,m]$)的潜在关联数量均为n,$C_y$($y\in[1,n]$)的潜在关联数量均为p,则A→C二阶知识发现的潜在关联总数为$m^n$,而A→D三阶知识发现的潜在关联总数为$m^{np}$,即总的潜在关联数量呈指数型增长,基于非相关文献的三阶知识发现方法将会发现更多的潜在关联。

(2)Kostoff曾将时序因素引入非相关文献知识发现方法[7-8],对于非相关文献的三阶知识发现方法,也可引入时序方法。

可作如下假设:对于某一领域的4个不同主题a、b、c、d,设a、b、c、d为时序序列,a为基础研究,b为技术主题,c为技术体系,d为工程应用,这是一个研究从基础理论到工程应用的过程。

既可以从基础理论研究 a 开始,挖掘与之相关的技术主题 b,通过这一技术主题构建一个技术体系 c,进而寻找这一技术体系与工程应用 d 的结合点;也可以从 d 的实际需求出发,寻找能解决这一需求的 c,然后寻找构成 c 的 b,最终攻关可以找到为 b 提供理论支撑的 a。

也就是说基于非相关文献的三阶知识发现方法可以以研究目标和具体问题为导向,为研究活动提供服务。

## 6 激光显示领域的实证

本文在 CNKI 中文数据库内,以"激光显示"为初始主题,进行了一次开放式的基于非相关文献的三阶知识发现方法研究的实验。实验数据的选取时间为 1980—2010 年的 CNKI 核心期刊论文。

### 6.1 实验流程

(1)A→B 过程

从"激光显示"主题为初始点,检索得到 26 篇文献,形成文献集 A,得到 A 集中的关键词共有 43 个。按出现频次排序、过滤,形成包含有 9 个关键词的中间主题 B(见表 2),构建文献集 B,共 101 篇文献。

**表 2 中间主题词 B**

| 主题 | 频次 |
|---|---|
| 激光散斑 | 2 |
| 振镜扫描 | 2 |
| 色域转换 | 2 |
| 超声光栅 | 1 |
| 电磁打点计时器 | 1 |
| 面阵空间光调制器 | 1 |
| 二维扫描转镜 | 1 |
| 全固态激光器 | 1 |
| 现场可编程门阵列 | 1 |

(2)B→C 过程

文献集 B 的 101 篇文献中包含有关键词 290 个,按出现频次排序、过滤,并通过人工选择得到 18 个关键词,形成中间主题 C(见表 3),构建文献集 C,共 347 篇文献。

**表 3 中间关联词 C(部分)**

| 主题 | 频次 |
|---|---|
| KTP 晶体 | 2 |
| 热透镜效应补偿 | 2 |
| 数字锁相环 | 1 |
| 移位寄存器 | 1 |
| 全固态腔内倍频激光器 | 1 |
| 端面泵浦 | 1 |
| 掺 Nd3 + 离子激光晶体 | 1 |
| Mie 氏散射 | 1 |
| Nd: YAG 激光器 | 1 |
| 量化误差 | 1 |

(3)C→D 过程

文献集 C 的 347 篇文献中包含有 817 个关键词。按出现频次排序、过滤,得到 768 个关键词,形成目标主题 D(见表 4)。

其中,A 与中间关联词 C、B 与目标词 D、A 与目标词 D 均无直接共现关系。

**表 4 目标主题 D(部分)**

| 主题 | 频次 |
|---|---|
| 相位噪声 | 22 |
| 光纤环形腔 | 10 |
| 光纤环形腔激光器 | 5 |
| 可变电容 | 5 |
| 双折射 | 5 |
| 掺铒光纤激光器 | 4 |
| 光纤光栅 | 4 |
| 视频水印 | 4 |
| 振荡调谐曲线 | 4 |
| 自启动 | 4 |

### 6.2 潜在关联的建立

通过多次尝试,本文尝试挖掘了一条潜在关联发现:

(1)A→B 的阶段

起始主题 A 为"激光显示",在上述文献集

A 中选取文献《激光显示中的色域转换系统》，其包含有关键字“激光显示、色域转换、现场可编程门阵列”。选取“色域转换”为中间关联词 B。

（2）B→C 的阶段

以“色域转换”为中间关联词 B，在文献集 B 中寻找与“色域转换”相关的研究，得到一篇文献《显示量化误差对色域转换的影响》，其包含有关键字“彩色显示、量化误差、色域转换”，进一步得到中间关联词 C 为“量化误差”。

（3）C→D 的阶段

以“量化误差”为中间关联词 C，在文献集 C 中寻找与“量化误差”相关的研究，得到文献 31 篇，形成目标词集 D，包含有关键字 87 个，即发现了 87 个 A→D 的潜在关联。

通过阅读与“量化误差”相关的 31 篇文献，寻找有意义的潜在关联。如文献《减小纯位相型计算全息图量化误差的一种新编码方法》，该文献“针对纯位相型计算全息图量化误差问题，提出了一种新的编码方法。其主要理论根据基于在复平面上可将一复矢量分解成任意两个复矢量之和”[9]。其包含有关键字“计算全息图、衍射光学器件、纯位相、量化误差、编码”，根据其内容，确定目标主题 D 为“编码”（基于复矢量法理论的新编码方法）。

A“激光显示”与 C“量化误差”、D“编码”，B“色域转换”与 D“编码”均无直接共现。

（4）建立关联

整个流程可概括为：A（激光显示）→B（色域转换）→C（减小量化误差）→D（基于复矢量法理论的新编码方法）。反向解读，即基于复矢量法，提出了一种新的编码方法减小了量化误差，进而提高了色域转换系统的性能，而色域转换系统是激光显示中的一项关键技术，这样便构成了一条“从实际需求（激光显示）出发，寻找这一应用的技术系统（色域转换系统），然后寻找提高这一系统性能的方法（减小量化误差），最终发现可以为该方法提供理论支撑的理论（复矢量法理论）”的潜在关联链条。

这也验证了上文提出的非相关文献的三阶知识发现方法可以“以研究目标和具体问题为导向，为研究活动提供服务”。

### 6.3 结果分析

本次实验主要是探讨验证基于非相关文献的三阶知识发现模式，因此在排序过滤算法中选取了简单的词频排序过滤，虽然会对发现效率和结果产生影响，但是仍然证明了以下几点：

（1）本实验证明，基于非相关文献的三阶知识发现方法可以发现大量潜在关联；中间关联 C 相对于中间关联 B 拥有更为庞大的中间词表，因此产生了更多的潜在关联。

（2）基于非相关文献的三阶知识发现方法可以以研究目标和具体问题为导向，为研究活动提供服务。可以说新方法实现了基于非相关文献的知识发现方法本身以“发现”的内涵为真正目的[8]。

## 7 结论及展望

（1）初步试验证明，基于非相关文献的三阶知识发现方法可以发现大量二阶知识发现方法无法发现的潜在关联，说明基于非相关文献的三阶知识发现方法值得深入探讨和研究。

（2）在本文的实验中仅对关键词进行了提取。在未来的研究中可将信息抽取范围扩大至文摘、全记录等。

（3）因为基于非相关文献的三阶知识发现方法中，生成两个中间集，具有三阶的传递关系，因此对 B 和 C 文献集的概念进行排序过滤的算法就会对文献集 C 和 D 生成叠加放大的影响。未来将对目前国内外非相关知识发现研究与实践中所应用的中间集处理方法进行比较分析，将根据基于非相关文献的三阶知识发现方法的不同发现模式的特点，选择适用于不同模式的各种排序过滤算法组合，并分析不同算法的组合对中间集及目标集产生的影响。

（4）在未来的研究中，还将对（下转第 69 页）

分析[J]. 图书馆工作与研究,2010(3):14 - 18.
[3] 叶鹰. 图书情报学前沿研究领域选评[J]. 中国图书馆学报,2008(4):63 - 69.
[4] 李长玲,翟雪梅. 基于硕士论文的我国图书馆学与情报学研究热点分析[J]. 情报科学,2008(7):1056 - 1060.
[5] 夏立新,金晶. 从 Google 网络图书馆计划的成功启动看图书馆数字化发展[J]. 情报科学,2009(4):485 - 488,492.
[6] 乔晓东,梁冰,李颖. 从 NSTL 战略定位到最新进展及未来发展规划[J]. 数字图书馆论坛,2010(10):11 - 17.
[7] 蒲筱哥. 我国特色数据库建设研究论文的统计分析[J]. 数字图书馆论坛,2009(9):53 - 56,65.
[8] 夏立新,韩永青,邓胜利. 基于知识供应链的图书情报机构知识服务模型研究[J]. 中国图书馆学报,2008(2):60 - 64,72.
[9] 杨宗英,郑巧英. 未来数字图书馆的发展方向之一——服务主导型数字图书馆[J]. 数字图书馆论坛,2004(5):20 - 23.
[10] 张正禄. 我国图书情报界云计算研究述评[J]. 国家图书馆学刊,2010(3):73 - 76,96.
[11] 刘炜. 图书馆需要一朵怎样的"云"[J]. 大学图书馆学报,2009(4):2 - 6.
[12] 孙卫. 图书馆在云时代的思考[J]. 数字图书馆论坛,2008(9):35 - 41.

**苏新宁**　南京大学信息管理系教授、博士生导师。通讯地址:南京市鼓楼区南京大学信息管理系。邮编:210093。
**夏立新**　华中师范大学信息管理系教授、博士生导师。通讯地址:武汉华中师范大学信息管理系。邮编:430079。

(收稿日期:2010-12-16)

(上接第 26 页)比分析不同阈值的选取对中间集产生的影响。将一种相关性算法的组合应用在基于非相关文献的三阶知识发现的各个排序过滤步骤中,在排序时尝试不同的阈值设置组合,对比分析不同阈值的选取对中间集产生的影响。

(5)正如本文实验中所列举的潜在关联链条所示,在基于非相关文献的三阶知识发现方法服务于科研活动这一目的上,还可以进一步挖掘该方法对于挖掘、创建从实际需求出发的创新链条的应用,在未来的研究中可作进一步研究和探讨。

**参考文献:**

[1] Swanson D R. Undiscovered public knowledge[J]. Library Quarterly, 1986(56): 103-118.
[2] Swanon D R. Fish oil, Raynaud's Syndrome, and undiscovered public knowledge[J]. Perspect Biol. Med,1986, 30:7-18.
[3] 安新颖, 冷伏海. 基于非相关文献的知识发现原理研究[J]. 情报学报,2006(1):87-93.
[4] 张云秋,冷伏海. 非相关文献知识发现的关键技术研究[J]. 情报学报,2008(4):521-527.
[5] The oracle of bacon[OL]. [2010-09-20]. http://oracleofbacon. org/.
[6] 王小凡. 复杂网络理论及其应用[M]. 北京:清华大学出版社, 2006.
[7] Kostoff R N, Jantke K P. Stimulating Innovation [J]. Lecture Notes in Computer Science, 2001, 2226: 196-213.
[8] 张树良,冷伏海. 基于文献的知识发现的应用进展研究[J]. 情报学报,2006(6):700-712.
[9] 杨茂田,丁剑平,王其和. 减小纯位相型计算全息图量化误差的一种新编码方法[J]. 南京大学学报(自然科学版), 2002(6): 842-849.

**李　勇**　中国科学院国家科学图书馆博士研究生。通讯地址:北京海淀区北四环西路 33 号中国科学院国家科学图书馆。邮编:100190。
**冷伏海**　中国科学院国家科学图书馆教授、博士生导师,情报研究部主任。通讯地址同上。
**王　林**　中国科学院国家科学图书馆博士研究生。通讯地址同上。

(收稿日期:2010-11-11)