基于文献关系模型的个性化文献管理方法①

王　炜，卢　罡，许南山

(北京化工大学 计算机科学与技术系，北京 100029)

摘　要：分析了国内外文献管理系统的现状。基于文献之间的属性特征，设计了文献关系模型，并且根据该模型定量分析文献之间的关联程度。在此基础上，设计并实现一个基于文献关系模型的个性化文献管理系统的原型系统。通过该系统可以对文献进行大量的个性化操作，方便用户使用。

关键词：文献管理系统；个性化；文献关系模型

Literature Relational Model-Based Personalized Literature Management System

WANG Wei, LU Gang, XU Nan-Shan

(Department of Computer Science and Technology, Beijing University of Chemical Technology, Beijing 100029, China)

Abstract: In this paper, we analyze the current situation of Literature Management. According to the attributes of literatures, a relational model of literatures is designed, and the interconnection between literatures is quantitatively analyzed. Based on the relational model, we design and implement a prototype of Personalized Literature Management System. Users of the system can manage their literatures in convenient and personalized way.

Key words: literature management system; relationship between documents; personalization; the relational model of documents

1　引言

计算机的普及以及 Internet 网络运用，促使人们更多的利用网络来获取信息，学术科研活动也在信息时代下产生巨大变革，Internet 已成为人们学习和交流信息的基本平台之一。目前，大量科技期刊文献都可以在网络中找到并下载。著名的此类网络文献[1]数据库有国内的重庆维普[2]、万方数据库、中国学术期刊全文数据库[3]等，国外的有 EBSCOhost[4]、Wiley InterScience[5]等，用户只需要登录到此类数据库就可以方便的查阅并下载文献信息。查阅和管理文献对研究工作具有重大意义，然而传统的文献管理是一件费时费力的工作，特别是当文献数量积累到一定程度后，仅仅靠大脑记忆很辛苦，也不可靠。传统的文献查阅和管理方式越来越无法满足要求，用户希望拥有便捷、高效、个性的管理工具来管理个人参考文献。在这种情况下，文献管理系统孕育而生。

文献管理系统又叫书目管理系统。文献管理系统就是一种让用户对各类型文献信息进行收集、管理、检索以及按照不同格式进行输入输出的工具[6]。

国外文献管理系统常见的有汤森路透公司的 Endnote[7]、Referencemanager[8]、ProCite[9]以及基于网络的 refworks[10]。其中 EndNote 是最受欢迎使用最多的软件，Reference Manager 提供网络功能可同时读写数据库，ProCite 提供弹性群组参考及可建立主题书目，refworks 让国内用户可以使用清华大学服务器来存储数据，不占用电脑空间和资源，用户可以随时随地访问个人文献书目数据库。

在中文文献管理系统中，有 NoteExpress[11]、PowerRef[12]、医学文献王[13]等优秀软件，其中 NoteExpress 是目前较好的中文文献管理系统，它将我们从易错、低效、繁琐的引文标注、参考文献编排工作中解脱出来，而且其检索便捷、功能强大、服务周到，将大大提高研究和写作论文的效率。PowerRef 融合单机资料管理与 Internet 数据查寻等功能，旨在帮助

① 收稿时间:2010-11-14;收到修改稿时间:2010-12-15

收集和管理文献资料，并以规范的格式把文献引用和注释输出到文字处理软件中，为撰写规范学士、硕士和博士学位论文以及其它各种类型学术科研论著与研究报告服务。医学文献王主要为医学科研人员建立和管理医学文献信息而设计。其功能包括从各种不同来源医学信息中汲取数据，建立个人文库，并将转换、浏览、编辑、编排参考文献和文献二次检索为一体。

随着 Internet 网络和计算机技术的不断发展，用户对文献管理系统功能的需求也越来越多。鉴于目前已经趋于成熟化的文献管理系统，今后以下两个方面是文献管理系统的发展方向：

（1）个性化方面：更友好的界面信息，更为符合个人用户的操作平台，使用户更容易上手学习。目前很多文献管理系统都需要用户花一定时间去了解使用方法和操作过程，未能达到“一看就会”的境地，这就给用户造成许多使用上的困难。同时，也使系统潜在用户望而却步。

（2）检索方法方面：趋于采用更为贴近用户习惯的检索方式，陈旧的检索方式已经使拥有大量文献资料的用户较为困难地快速寻找所需文献或记录。创建符合用户习惯的检索方式，大大减少检索时间，提高检索的准确度，将是未来文献管理系统发展方向。

文献管理系统主要给用户提供一个管理文献的操作平台，这个操作过程是非常主观的。在日常文献管理中，管理对象是文献，若能基于文献之间的关系进行管理，更符合用户的思维习惯。揭示这种内在关系，能最大限度方便用户管理，让用户一目了然，这对用户而言将起到巨大的正面效用。基于此思想，本文提出基于文献属性的文献关系模型，并基于该模型初步实现了个性化文献管理原型系统。

2 文献关系模型

文献之间最重要的关系之一，就是文献的引用关系。这种引用关系可以反映出某研究内容的发展过程，也可以通过文献被引次数等情况，反映出文献重要程度。这里，我们通过定义文献繁衍深度，来反映文献之间的关联程度。

此外，基于文献属性，定义了文献关联距离。类比文献之间各种属性，把相同属性作为量化文献之间关联距离的依据，结合文献繁衍深度，进行文献关联距离计算。下面分别介绍文献繁衍深度模型和文献关联距离模型。

2.1 文献繁衍深度关系模型

定义 1. 一篇文献 l 为一 m 维的向量：

$$l=(a_1,\ a_2,\cdots,\ a_{(m-1)},\ a_m),$$

其中，$(a_{i1},\ a_{i2},\cdots,\ a_{i(m-1)},\ a_{im})$ 为文献 l_i 的 m 个属性分量。

定义 2. 对于一个文献的集合 $L=\{l_1,l_2,\ldots,l_n\}$，可用一个 $n\times m$ 矩阵 $\boldsymbol{M}_L$ 表示如下：

$$M_L=\begin{bmatrix}l_1\\l_2\\\vdots\\l_n\end{bmatrix}=\begin{bmatrix}a_{11},\ a_{12},\cdots a_{1(m-1)},\ a_{1m}\\a_{21},\ a_{22},\cdots a_{2(m-1)},a_{2m}\\\vdots\\a_{(n-1)1},\ a_{(n-1)2},\cdots a_{(n-1)(m-1)},\ a_{(n-1)\,m}\\a_{n1},\ a_{n2},\cdots a_{n\,(m-1)},\ a_{nm}\end{bmatrix}$$

其中，每一行为一篇文献的属性组，每一列为文献的一种属性。

定义 3. 定义操作 $Ref(l)$，以获取文献 l 的所有参考文献集合。则有 $Ref(l)=\{r_1,r_2,\ldots,r_{n'}\}$，其中 $r_i,i=1,2,\ldots,n'$ 为文献 l 的参考文献。

定义 4. 定义操作 $Cite(l)$，以获取所有引用文献的文献集合。则有 $Cite(l)=\{c_1,c_2,\ldots,c_{n''}\}$，其中 $c_i,i=1,2,\ldots,n''$ 为引用文献 l 的文献，称为“引用文献”。

定义 5. 设文献 l_1、l_2 和 l_3，若有 $l_2\in Ref(l_1)$，且 $l_3\in Ref(l_2)$，则定义 $l_3\in Ref^2(l_1)$，表示 l_3 为 l_1 的参考文献的参考文献，称 l_3 为 l_1 的“二代参考文献”。同理，$l_p\in Ref^n(l_q)$ 表示 l_p 为 l_q 的“n 代参考文献”。

定义 6. 设文献 l_1、l_2 和 l_3，若有 $l_2\in Cite(l_1)$，且 $l_3\in Cite(l_2)$，则定义 $l_3\in Cite^2(l_1)$，表示 l_3 为 l_1 的引用文献的引用文献，称 l_3 为 l_1 的“二代引用文献”。同理，$l_p\in Cite^n(l_q)$ 表示 l_p 为 l_q 的“n 代引用文献”。

定义 7.进一步地，若有 $l_2\in Ref(l_1)$，$l_3\in Ref(l_2)$，…，$l_n\in Ref(l_{n-1})$，即 $l_n\in Ref^{n-1}(l_1)$，则称序列 $l_1,l_2,\ldots,l_n$ 为“上行繁衍关系序列”。同理，若有 $l_2\in Cite(l_1)$，$l_3\in Cite(l_2)$，…，$l_n\in Cite(l_{n-1})$，即 $l_n\in Cite^{n-1}(l_1)$，则称序列为 $l_n,l_{n-1},\ldots,l_2,l_1$ “下行繁衍关系序列”。上行繁衍关系序列和下行繁衍关系序列统称为文献的“繁衍序列”。两个文献和之间的繁衍序列可能有多个，长度最短的繁衍序列的长度称为文献和之间的繁衍深度，记作：N_{pq}，$N_{pq}\in[0,+\infty]$ 且 $N_{pq}\in N$。

定义 8. 文献对其自身的繁衍深度定义为 0；两篇无繁衍关系的文献其繁衍深度定义为 $+\infty$。

2.2 文献关联距离模型

根据定义 1，可以定义文献的属性分量。由于文献的属性分量较多，篇幅有限，这里只列出一些属性分量：作者名、论文类型、题目、字数、重要程度、中文摘要、英文摘要、中图分类号、文献标识码、数字对象唯一标识符、文章编号、关键字、作者通讯方式、作者所在单位、作者国籍、所属出版社、出版年份、文献语言、文献国别等，其中重要程度属性由用户自行定义。重要程度分为十个级别，采用 10 分制表示。1 分代表极不重要，10 分代表极重要，2 到 9 分为中间渐变过程分值。十级分制主要是便于用户标识文献的不同重要性。下面以这些属性分量为例说明文献关联距离模型。

定义 9　设二元组集合 K=$K_1 \cup K_2$，其中 K 为文献属性分量集合，K_1 为文献属性分量中可与其他文献相同属性分量构成关联关系的属性分量集合，K_2 为不可构成关联关系的属性分量集合。

另外 K_1 包含：作者名、论文类型、重要程度、作者通讯方式、作者所在单位、作者国籍等。K_2 包含：图分类号、文献标识码、数字对象唯一标识符、文章编号、字数等。

在实际情况中，文献某一属性分量值并不唯一。如一篇文献可以有多个作者，那么此文献作者属性分量值就有多个。根据这种情况我们再次对集合 K_1 进行划分。

定义 10. 设二元组集合 G=$G_1 \cup G_2$，其中 G_1 和 G_2 的元组 K_1 的子集，其中：

G_1 为单值型属性分量集合，G_2 为集合型属性分量集合，且 $G_1 \cup G_2 = K_1$，$G_1 \cap G_2 = \Phi$。

由以上定义可知：

G_1 包括所属出版社、所属期刊、所属会议、重要程度等单值型属性分量。G_2 包括关键字、指导老师、作者名等集合型属性分量。

定义 11. 在文献属性分量子集 G_1 中，两篇文献之间某属性分量值相同，那么认为这两篇文献的此属性分量有关联。设 $a \in G_1$ 为文献 l_p、l_q 的属性分量，$l_p(a)$、$l_q(a)$ 为文献 a 属性分量值。我们把文献 l_p、l_q 对 a 属性的关联关系量化为 a 属性距离 $X_{pq(a)}$。

$$X_{pq(a)} = \begin{cases} 0 & l_p(a) = l_q(a), \\ 1 & l_p(a) \neq l_q(a). \end{cases}$$

对于文献的出版年份这一属性，其属性距离作如下定义：

定义 12. 假设 $y \in G_1$ 为文献 l_p，l_q 的出版年份属性分量，$l_p(y)$、$l_q(y)$ 为文献 y 属性分量值。那么根据逻辑斯蒂变换：$f(x) = ln\ \frac{x}{1-x}$，将 y 属性距离 $X_{pq_{(y)}}$ 标准化,定义为：

$$X_{pq_{(y)}} = 2*\left(1 - \frac{1}{1+e^{|l_p(y)-l_q(y)|}}\right) - 1$$

定义 13. 假设 $s \in G_1$ 为文献 l_p，l_q 的重要程度属性分量，$l_p(s)$、$l_q(s)$ 为文献 s 属性分量值。那么将 s 属性距离 $X_{pq(s)}$ 标准化，定义为：

$$X_{pq(s)} = 2*\left(1 - \frac{1}{1+e^{|l_p(s)-l_q(s)|}}\right) - 1$$

定义 14. 对于文献属性分量子集 G_2 中的属性分量，集合型属性分量间的距离计算方法定义如下：

设 $b \in G_2$ 为文献 l_p、l_q 的属性分量，文献 l_p 中 b 属性分量有 m 个属性分量值 $l_p(b)_i\ (i=1,2,\cdots,m)$，文献 l_q 中 b 属性分量有 n 个属性分量值 $l_q(b)_j\ (j=1,2,\cdots n)$。记 $l_p(b) \cap l_q(b)$ 中元素个数为 r。则文献 l_p、l_q 的 b 属性分量的属性距离 $X_{pq(b)}$ 定义为：

$$X_{pq(b)} = 1 - \frac{r}{m+n-r}$$

同时，为了方便文献繁衍深度关系模型与文献关联距离模型的结合，我们给出文献之间繁衍深度距离概念。

定义 15. 文献 l_p、l_q 之间的繁衍深度距离 $X_{N_{pq}}$ 定义如下：$X_{N_{pq}} = 2*\left(1 - \frac{1}{1+e^{N_{pq}}}\right) - 1$。

基于以上定义，由加权欧氏距离定义文献之间的关联距离。

定义 16. 文献 p 和文献 q 的关联距离 X_{pq} 定义如下：

$$X_{pq} = \sqrt{\sum_{i=1}^{m}\alpha_i\left(X_{pq(a)}\right)^2 + \sum_{j=1}^{n}\alpha_j\left(X_{pq(b)}\right)^2 + \alpha_{N_{pq}}\left(X_{N_{pq}}\right)^2 + \alpha_y\left(X_{pq_{(y)}}\right)^2 + \alpha_s\left(X_{pq_{(s)}}\right)^2}$$

其中，m 为文献属性分量子集 G_1 中元素个数，$a \in G_1$。n 为文献属性分量子 G_2 集中元素个数，$b \in G_2$。α 为权值系数，$\{\alpha \mid 0 \le \alpha \le 10,\ \alpha \in R\}$。令 α_z 为各项权值的总和，$\alpha_z = 10$。

2.3 文献关联距离模型实例

下面举三篇文献为实例说明文献关联距离模型：

2.3.1 材料文献(如下表)：

表 1 三篇材料文献详细信息表

	文献 1	文献 2	文献 3
题录类型	期刊文章	期刊文章	期刊文章
作者	赵伟艇,史玉珍	闫磊,张国圆,周海燕	张晓芬
年份	2010	2007	2000
繁衍关系	上行繁衍文献 2	下行繁衍文献 1	无繁衍关系
标题	基于 802.11i 的无线局域网安全加密技术研究	无线局域网的安全性研究	EBSCO 网络数据库综合评价
期刊名	计算机工程与设计	福建电脑	现代图书情报技术
期刊类型	计算机类	计算机类	图书情报类
关键词	无线局域网 ,IEEE802.11i, 安全,加密,临时密钥完整性,协议,	无线局域网, 安全,攻击,	网络数据库,质量评价,EBSCO 公司,综合评价,检索,
作者地址	平顶山学院,河南平顶山	安徽省中国矿业大学信电学院	北京科技大学图书馆
作者国籍	中国	中国	中国
文献语言	汉语	汉语	汉语
期刊所在国	中国	中国	中国
期刊语言	汉语	汉语	汉语
重要程度	7	9	3

2.3.2 计算三篇文献的关联距离

为体现不同权值对文献关联距离的影响，令文献繁衍深度权值系数为 9（$\alpha_{N_{pq}}=9$），其余属性分量权值系数为 0.1。根据定义 16，计算各文献之间的关联距离为：

①文献 1 与文献 2 的关联距离（$e=2.718$）：

$$X_{12}=\sqrt{0+0.1+0.1*\left(X_{12_{(y)}}\right)^2+9*\left(X_{N_{12}}\right)^2+0.1+0.1+0+0.1*\left(1-\frac{2}{3+6-2}\right)^2+0.1+0+0.1*\left(X_{12_{(s)}}\right)^2}$$

其中，$X_{12_{(y)}}=2*\left(1-\frac{1}{1+e^3}\right)-1=0.905$，

$$X_{N_{12}}=2*\left(1-\frac{1}{1+e^1}\right)-1=0.462,$$

$$X_{12_{(s)}}=2*\left(1-\frac{1}{1+e^2}\right)-1=0.762。$$

原式$=\sqrt{0.4+0.1*(0.905)^2+9*(0.462)^2+0.051+0.1*(0.762)^2}=1.585$。

②文献 1 与文献 3 的关联距离：

$$X_{13}=\sqrt{0+0.1+0.1*\left(X_{13_{(y)}}\right)^2+9*\left(X_{N_{13}}\right)^2+0.1+0.1+0.1+0.1+0.1+0+0.1*\left(X_{13_{(s)}}\right)^2}$$

其中，

$$X_{N_{13}}=2*\left(1-\frac{1}{1+e^{+\infty}}\right)-1=1,$$

$$X_{13_{(s)}}=2*\left(1-\frac{1}{1+e^4}\right)-1=0.964。$$

原式$=\sqrt{0.6+0.1*1+9*1+0.1*(0.964)^2}=\sqrt{9.7+0.1*(0.964)^2}=3.129$。

③文献 2 与文献 3 的关联距离：

$$X_{23}=\sqrt{0+0.1+0.1*\left(X_{23_{(y)}}\right)^2+9*\left(X_{N_{23}}\right)^2+0.1+0.1+0.1+0.1+0.1+0+0.1*\left(X_{23_{(s)}}\right)^2}$$

其中，

$$X_{23_{(y)}}=2*\left(1-\frac{1}{1+e^7}\right)-1=0.998,$$

$$X_{N_{23}}=2*\left(1-\frac{1}{1+e^{+\infty}}\right)-1=1,$$

$$X_{23_{(s)}}=2*\left(1-\frac{1}{1+e^6}\right)-1=0.995。$$

原式$=\sqrt{0.6+0.1*(0.998)^2+9+0.1*(0.995)^2}=3.130$。

2.3.3 结论

在两类计算中，文献 1 与文献 2 的关联距离 X_{12} 取得最小值，关联程度最大。文献 2 与文献 3 的关联距离 X_{23} 最大，其关联程度最小。从现实情况看，文献 1 与文献 2 的关联程度确实相比各自与文献 3 的关联程度都要大，文献关联距离模型的计算结果符合实际情况。证明文献关联距离模型是正确有效的。

3 系统设计与实现

基于以上提出的文献关系模型，本文设计了文献管理原型系统。下面介绍系统设计框架和实现的功能。

3.1 总体框架设计

本系统分为三大模块，如图 1 所示，分别是：常规管理，各类检索和用户定义模块。各个模块又分为若干相应功能。常规管理功能使用户对文献进行增、删、改等基本操作。各类检索模块为用户提供多种检索模式，同时使用户取得对文献的最大筛选能力。用

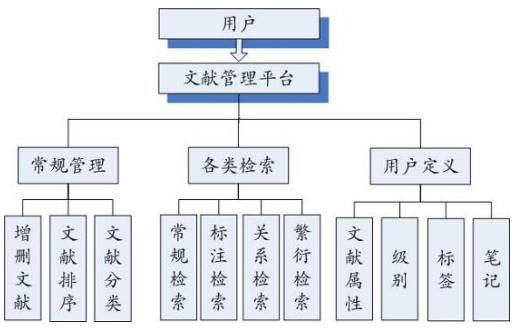

图 1 文献管理系统功能模块图

户自定义模块让用户非常个性化的通过标签、笔记和文献级别等属性来设定自己的文献，使整个系统的管理和操作更具人性化。

3.2 软件主要功能介绍

（1）繁衍检索：繁衍检索的数学推导模型在前文已经介绍。许多科研人员和查阅文献用户，需要对某一研究领域进行深入的了解，在阅读大量文献时，往往会检索某篇文献的引用或参考文献作为阅读文献，通过阅读暗含时间序列的繁衍关系文献，用户不仅仅可以了解到目前国际国内这一领域的前沿和热点，同时还可以了解其发展历史。繁衍检索功能正是在这种需求背景下构思设计，通过找出文献与其参考文献的“繁衍关系序列”，可以极大的方便用户了解研究课题演变过程，深入课题方向，加快科学研究效率。

（2）关系检索：通过上一节定义的两个数学模型，我们找到了量化计算两篇文献关联程度的方法，使用这种方法可以计算出任意两篇文献的关联距离，根据关联距离的大小可以对文献进行排序和检索等操作。关系检索针对文献各种属性进行拟合比较，通过这种方式构建一张文献关系网络，用户可利用此关系网络实现个性化检索。

（3）标注检索：通常用户对文献的标注（标签、笔记和文献重要等级），来检索文献。较多情况下，查阅文献以后，用户还需要重新查阅某一篇已经阅读过的文献或者已经标记的文献，由于文献数量巨大，显然不能通过把所有文献再看一篇的方式来查找。解决办法是在用户阅读文献中就对文献做注释，这些注释可以记录作者的想法和对文献的批注等。只要用户对文献有所标注，那么通过检索标签、笔记和自己判定的文献等级等几种手段，就能很快地找到目标文献。

4 结论

本文通过对国内外文献管理系统的调研，分析其现状与发展趋势，针对文献之间的关系设计了文献繁衍关系模型和文献关联距离模型。利用这些关系模型，量化了文献之间的关联程度，并且实现了个性化文献管理原型系统。同时，实现文献管理系统原型也证明了本文所提出的两个文献关系模型的可行性和实用性。这对文献管理系统的研究和发开具有一定指导意义。

参考文献

1 崔秀文.论网络文献信息的开发、利用与共享.科技与管理,2003,5(3):134-136.

2 张成全.有效使用维普数据库的实践与探索.浙江交通职业技术学院学报,2002.87-88.

3 孟秀华,陈培民,张敏聪.内蒙古农业大学图书馆主要文献数据库及其检索方法.内蒙古农业科技,2005.37-39.

4 张晓芬.EBSCO 网络数据库综合评价.北京科技大学图书馆, 2000.72-74.

5 Wileynter Science. http://www3.interscience.wiley.com/cgi-bin/home.

6 梁春燕,李晓霞,聂峰光,郭力,杨章远.基于B/S模式的科技文献管理系统的策略和初步实现.现代图书情报技术,2004.24-28.

7 Agrawal A. EndNote1-2-3Easy!: Reference Management for the Professional. Springer Publishing Company, Incorporated, 2009:296.

8 张爱红.常用参考文献管理软件的介绍与特点比较.中国科教创新导刊,2008.128-128.

9 Canos JH. A bibliography manager for Microsoft Word. ACM 2 Penn Plaza,2000:27-30.

10 张苏,张建.个人文献管理及参考文献创建工具——RefWorks 使用技巧.图书馆杂志,2007(5).

11 谢奇,李立立,关中玉,高桐.NoteExpress——中文科技文献管理的绝佳助手.科技文献信息管理,2007,21(3):18-21.

11 谢奇,蒋若冰,关中玉.实用的中文参考文献管理软件——PowerRef.科技文献信息管理,2008,22(2):1-5.

12 谢奇,李立立,蒋若冰.科研工作者必备的文献管理软件——《医学文献王》.中华医学图书情报杂志,2008,17(1):63-65.