情报学报 ISSN 1000-0135
第29卷第6期1066-1073，2010年12月

JOURNAL OF THE CHINA SOCIETY FOR SCIENTIFIC AND TECHNICAL INFORMATION ISSN 1000-0135
Vol.29 No.6,1066-1073 December 2010

doi:10.3772/j.issn.1000-0135.2010.06.015

基于聚类分析的学科交叉研究[1)]

魏建香[1,2] 孙越泓[3] 苏新宁[1]

（1. 南京大学信息管理系，南京 210093；2. 南京人口管理干部学院信息科学系，南京 210042；
3. 南京师范大学数学与计算机学院，南京 210097）

摘要 聚类分析是数据挖掘中的一项重要技术，通过聚类可以发现隐藏在海量数据背后的知识。本文提出了一种通过文献数据聚类分析来研究学科交叉的方法。首先提出了一种基于摘要词与关键词加权的相似度模型，使得文献之间的相似度更加精确。利用FCM算法对2005年CSSCI文献数据库中图书情报学的文献数据进行聚类，通过建立学科原子特征词的学科交叉表统计出图书馆学、情报学和文献学三个学科的研究热点及交叉点，以及图书情报学新的学科增长点，并对分析结果进行了检验，结果表明本文所提出的方法是科学的、切实可行的。

关键词 聚类分析 学科交叉 相似度 FCM

Research on Interdisciplinarity Based on Document Clustering Analysis

Wei Jianxiang[1,2], Sun Yuehong[3] and Su Xinning[1]

(1. *Department of Information Management, Nanjing University, Nanjing* 210093;
2. *Department of Information Science, Nanjing College for Population Programme Management, Nanjing* 210042;
3. *School of mathematics Science, Nanjing Normal University, Nanjing* 210097)

Abstract Clustering analysis is an important research field of data mining, through which we can find hidden knowledge behind mass data. This paper presents a method to study interdisciplines by document clustering analysis. At first, we design a similarity computing model based on the weighted contribution of summary words and keywords, which can make similarities between documents more accurate. Then, we choose the data of 2005' Library and Information Science from CSSCI literature database and apply the classic FCM algorithm to implement data clustering. Finally, we analyze the clustering results through statistic analysis by creating the interdisciplinary table based on disciplinary atomic words of library science, information science, and philology. The results include research hotspots and cross-point, and new discipline growth points of the three disciplines. By testing, the proposed method of this paper proves scientific and feasible.

Keywords clustering analysis, interdiscipline, similarity, FCM

1 引 言

科学研究需要创新。科学技术的进步为每个学科带来新的发展机遇的同时，也带来更为严峻的挑战。目前，文、理、工、管等学科之间相互渗透、交叉、融合已经成为一种潮流和趋势，其深度和广度正在进一步深化。学科通过引入、吸收、整合其他学科的

收稿日期：2009年10月28日

作者简介：**魏建香**，男，1972年生，副教授，南京大学信息管理系博士生，南京人口管理干部学院信息科学系副主任，主要研究方向：人工智能、数据挖掘研究等。E-mail:jxwei@foxmail.com。**孙越泓**，女，1972年生，副教授，博士研究生，主要研究方向：人工智能、图像处理等；**苏新宁**，男，1955年生，南京大学信息管理系博士生导师，教授，主要研究方向：信息处理与检索、知识管理、引文分析等。

1) 国家社科基金青年自选项目(09CTQ022)；江苏省"六大人才高峰"第六批资助项目(09-E-016)；教育部人文社会科学重点研究基地2008年度重大项目(08JJD870225)。

理论、方法和技术来促进学科自身的发展是学科创新最常见的途径。例如,计算机学科中的一个非常热门和活跃的研究分支-人工智能中的许多优化算法:免疫算法、遗传算法、蚁群算法等都是借鉴生物学理论提出的。因此研究学科交叉可以反映学科研究的热点和发展趋势,为研究者所从事的学科研究方向提供决策支持。作为一名科研工作者,必须清醒地认识到自己所从事的学科在机遇和挑战面前,能否选择正确的研究方向,能否走在学术研究的前沿,能否为学科的发展做出贡献,这是每一个科研工作者必须深入思考的问题。目前,许多文献数据库已经建成并投入使用,如万方、中国知网、维普等数据库,积累了海量学术文献,为研究学科交叉提供了数据保障。聚类技术是数据挖掘中的一项重要技术,文献聚类就是依据文献之间的相似度按照一定的算法准则,将文献进行动态模糊的分类,通过聚类可以挖掘数据背后隐性知识,从而揭示学科交叉。中国科学院院长路甬祥院士指出:"学科交叉点往往就是科学新的生长点,新的科学前沿,这里最有可能产生重大的科学突破,使科学发生革命性的变化。同时,交叉科学是综合性、跨学科的产物,因而有利于解决人类面临的重大复杂科学问题、社会问题和全球性问题。在新时期里,中国需要加速发展科学和技术,其中要大力地提倡学科交叉、注重交叉科学的发展。因而,提出并解决交叉科学难题就具有重大的意义"[1]。从中我们清楚地看到交叉学科对于社会和经济的发展具有举足轻重的意义,因而学科交叉研究也同样具有十分重要的意义:通过对文献数据的聚类分析,挖掘学科交叉点,使研究者了解本学科目前的研究现状,如学科发展前沿与热点问题等,以提高研究者的创新意识和创新动力,为科学研究提供决策支持;为管理者和研究机构提供决策支持,如交叉学科的政策支持、研究经费投入、人才培养方向等;通过学科交叉的比较,使学科本身获得动力,提升学科竞争力,使学科能更好地适应社会和经济的发展,更好地服务社会。本文正是在这种背景下,提出了一种基于文献聚类分析的学科交叉研究方法,并通过实例对提出的方法进行了分析与验证。

2 研究现状分析

国外一般把学科交叉称为"跨学科"。1970 年在法国召开了国际学术讨论会,出版了三卷本《跨学科—大学中的教学和研究问题》,这标志着跨学科学这门新学科的诞生。同年《跨学科综合杂志》、《跨学科历史杂志》创刊。1974 年《国际跨学科研究年鉴》问世。1976 年国际性的《交叉科学评论》在英国创刊,标志着研究进入了新阶段。1979 年美国出版"人文科学跨学科研究生计划"研讨会文集《高等教育中的跨学科》。同年国际跨学科学协会成立,相继在德、英、美、法召开会议。1981 年法国出版雷斯韦伯的《跨学科方法》,1986 年美国出版《跨学科分析和研究》,这些成果标志着跨学科学已达到新的水平。1990 年美国出版了克莱茵的《跨学科学—历史、理论和实践》,从多学科视角研究了跨学科基本理论和应用实践等。2000 年,加拿大出版的《实践中的跨学科学》更突出了跨学科的应用实践。

我国学科交叉研究萌生于 20 世纪 50 年代,到 80 年代进入全面展开的阶段。1985 年,钱学森、钱三强、钱伟长等在"全国首届交叉科学讨论会"上作了重要讲话。1986 年,跨学科学会试办的《交叉科学》杂志创刊,此后的研究论文增多,主要研究学科交叉的概念、产生缘由、历史、地位、功能等;90 年代前后,李光、任定成主编的《交叉科学导论》、刘仲林主编的《跨学科学导论》、徐飞的《科学交叉论》等学科交叉理论研究专著相继问世。1997 年以来,金吾伦的《跨学科研究引论》、刘仲林的《现代交叉科学》等专著陆续出版,代表了学科交叉研究的新成果[2]。

目前有关交叉学科的文献研究的主要特点是:①主要集中在学科交叉的概念、意义、现状分析及交叉模式的探讨[3~9];②注重理论探讨,缺乏数据支撑。有关文献聚类的学术论文的主要特点:①单纯从技术的角度研究聚类算法的设计与改进[10~15];②以文献关键词作为聚类的依据,不能充分地、科学地反映学科主题[16,17];③聚类结果无交叉,不能将文献划分至不同的类别。从中国知网文献数据库中,检索题名或关键词中含有"学科交叉"的文献共有 546 篇,基于文献聚类的学科交叉研究的文献是 0 篇,因此本文的研究具有一定的前沿性。

3 研究思路

3.1 研究框架

文献是反映学术研究的主要依据,虽然目前已经发表的学术论文都有确定的分类标准(按中图分类号进行标注),但一些文献可能属于多个学科领域,也是学科交叉的代表文献。事实上,正如程桂瑛在《对一些涉及新兴学科、交叉学科的文献分类之我

见》中提到的"随着科学技术的迅速发展,新兴学科、边缘学科、多元交叉学科方面的文献越来越多,使从事图书、资料分类工作的人员普遍感到有些文献很难归类"。文献所归属的类别是由所发表的期刊给定的,而交叉学科文献的学科归类通常情况下是不易确定的、较模糊的,这与聚类技术本身是一个动态模糊的过程是一致的,我们可以利用数据挖掘中的常用技术-聚类分析来实现学科交叉的研究。基于关键词来进行文献聚类是目前采用最普遍的方式,但由于文献关键词数量少及作者选择关键字的不规范和不科学等因素的影响使得关键词并不一定完全反映文献学科的主题,影响了聚类的效果,我们采用关键词与摘要词加权的方法以提高文献聚类质量。本文的研究是在海量文献数据的基础上,采用聚类分析的理论和方法,挖掘隐藏在数据背后的学科交叉知识,研究的整体框架如下:

(1)利用现有的文献数据库,采集图书馆学、情报学、文献学权威期刊的文献数据,经过抽取和提炼,建立学科特征词库,用于类别的学科特征识别;

(2)提取文献关键字和摘要词并进行降维和加权处理,建立文献向量空间模型矩阵,采用FCM聚类算法实现文献聚类;

(3)根据聚类结果,建立用于统计分析的学科交叉表;

(4)学科交叉结果的统计分析。

3.2 向量空间模型的构建

向量空间模型(VSM)是文献分类所使用的特征较为普遍采用的方法之一。在这个模型中,文献空间被看作是由一组正交词条向量组成的向量空间,每个文献表示为其中的一个范化特征向量:$V(d)=(t_1,w_1(d),\cdots,t_i,w_i(d),\cdots,t_n,w_n(d))$,其中 t_i,可以要求 t_i 是 d 中出现的关键短语,以提高文献内容的准确性。$W_i(d)$常被定义为 t_i 在 d 出现频率 $tf_i(d)$的函数,如 $w_i(d)=(tf_i(d))$,常用的函数有布尔函数:$\Phi=\begin{cases}1, tfi(d)\geqslant 1\\0, tfi(d)=0\end{cases}$;平方根函数:$\Phi=\sqrt{tfi(d)}$;对数函数:$\Phi=\log(tfi(d)+1)$;*TFIDF*函数:$\Phi=tfi(d)\times\log\left(\frac{N}{ni}\right)$等。本文采用了一种新的 $w_i(d)$的计算模型,具体步骤如下:

3.2.1 构建特征向量空间

文献聚类是在一个非常高的维度中进行的,而聚类算法的复杂度与数据维度是非线性关系。理论证明,随着维度的增加,计算的复杂度将呈现指数级的增长。我们通过对2005年CSSCI文献数据库中三大学科文献的统计发现,在3932篇文献中关键字的个数为14202,平均每篇文献的关键词个数为3.61个,互异的关键词达到6708个,利用普遍认可的VSM来计算文献相似度矩阵时,特征空间的维度很大,增加了计算的复杂性。因此,文献聚类的首要问题是要将数据进行降维。原子特征词是指从所有文献关键词中找出能够反映出学科特点的关键词中的原子部分。这样做可以将数组维度从一个非常高的维度降低到一个维度相对较低的空间。

原子特征词的确立是基于关键词的文本分类的非常关键的基础问题,词典中原子特征词的选取既要考虑这些词在文本集合中出现的统计特征,选取那些反映文本内容的原子特征词;又要做停用词表,去掉那些在特定语言中出现频率较高但含义虚泛的词,以降低特征空间的维数;同时还要考虑关键词的频幅限制,以防止因少数关键词在少数文本中频幅过高而造成的聚类中心的偏移影响。

3.2.2 关键词相似度计算模型

由于我们对关键词进行了抽取和加工,因此大部分的文献关键词与原子特征词并不完全相同,由

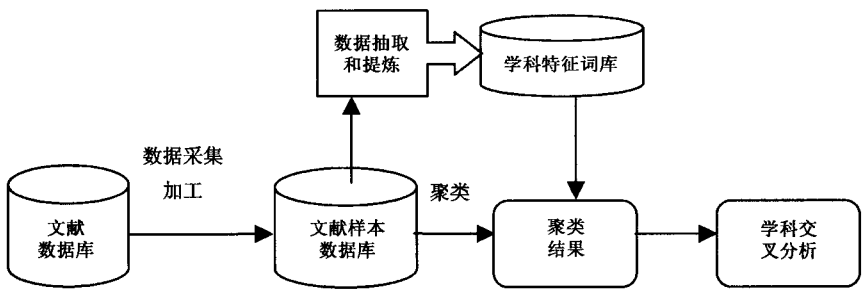

图1 学科交叉研究框架

于在传统的相似度计算模型中两者的相似度将是0,这会使得构建的文献空间向量矩阵是绝大部分元素出现0的稀疏矩阵。因此,必须考虑两种关键词之间的部分相似性[18]。

假设两个关键字 k_i 和 k_j,字符长度分别为 l_i 和 l_j,连续相同字符串长度为 l,则该两个关键字相似度定义为 $T(k_i,k_j)$:

$$S(k_i,k_j)=\begin{cases}\dfrac{l}{l_i+l_j-l} & \text{当 } l>=4\\ 0 & \text{当 } l<4\end{cases} \quad (1)$$

显然有 $T(k_i,k_j)\in[0,1]$。这个公式考虑了关键字之间的部分相似性,提高了相似度计算的精度。例如:两个关键字"公共图书馆"和"数字图书馆",在许多的文献聚类方法中将这两个关键字的相似度定义为0(即两者完全不同),这在某种程度上影响了文献相似度的精确度。利用我们给出的公式(1)计算结果为0.4286,能更加准确地表示两者的相似度。

3.2.3 关键词与摘要词相似度加权计算模型

每一篇文献一般有若干个关键词,通过关键词相似度计算模型中公式(1)计算所得的值也相应地有若干个。为了进一步提高文献相似度的精确度,我们结合关键词相似度计算模型和关键词的频次提出一种新的计算文献相似度的相似度加权计算模型:

假设第 i 篇文献 d_i 的关键词有 p 个,关键词集合定义为 $D=(k_1,k_2,\cdots,k_p)$,查找这些关键词在文献摘要中出现的频次,记为 $F(D)=(F(k_1),F(k_2),\cdots,F(k_p))$,通过关键词相似度计算公式(1)计算 D 中的关键词与第 j 个原子特征词 k_j 之间的相似度:$S(D)=(S(k_1,k_j),S(k_2,k_j),\cdots,S(k_p,k_j))$,则文献 d_i 在原子特征词 k_j 处的取值可定义如下:

$$w_{ij}=\max(S(k_1,k_j)\times(F(k_1)+1),S(k_2,k_j)\times(F(k_2)+1),\cdots,S(k_p,k_j)\times(F(k_p)+1)) \quad (2)$$

3.2.4 文献向量矩阵的构建

空间向量模型的目标是将文献数据表示为特征空间中的一个向量,所有文献向量构成文献向量矩阵。该矩阵以文献作为行,以原子特征词为列。假设 n 篇文献,m 个学科原子特征词,则文献的空间向量矩阵表示为:

$$\begin{bmatrix}w_{11} & w_{12} & \cdots & w_{1m}\\ w_{21} & w_{22} & \cdots & w_{2m}\\ \vdots & \vdots & & \vdots\\ w_{n1} & w_{n2} & \cdots & w_{nm}\end{bmatrix}$$

其中,文献在 m 维空间的特征分量 w_{ij}($i=1,2\cdots,n;j=1,2\cdots,m$)计算方式见公式(2),是通过每篇文献的关键词与原子特征词之间的相似度乘以该关键词在摘要中出现的频次加1,并取最大值得到。例如,第1篇文献中的3个关键词及在摘要中出现的频次分别为:数字型图书馆(3次);隐性知识(4次);图书馆工作(2次),第1维的原子特征词为"图书馆"。则这篇文献的 w_{11} 的计算方式为:首先利用相似度计算模型中公式(1)计算3个关键词与原子特征词"图书馆"的相似度值分别为0.5、0、0.6,然后计算加权后的值分别为:$0.5\times(3+1)=2.0$、$0\times(4+1)=0$ 及 $0.6\times(2+1)=1.8$,取最大值2.0作为 w_{11} 的值。

4 文献聚类的实现

为了研究学科交叉,我们选择了2005年CSSCI图情档文献数据,共计3932篇文献,来源于30种期刊。

(1)数据抽样:从3932篇文献中,通过系统抽样法,从中选取了800篇文献,其中情报学257篇,文献学120篇,图书馆学423篇。

(2)数据清洗:通过查询中知网(www.cnki.net),取出每一篇文献的摘要部分,并统计每一篇文献中关键词在摘要中出现的频次,去掉其中表达不规范的摘要,最终选取情报学、文献学、图书馆学三个学科文献共681篇,其中情报学207篇,文献学98篇,图书馆学376篇。

(3)抽取原子特征词:取出681篇文献中所有的关键词,从中人工选取108个原子特征词,如表1。通过编写程序,求出每一篇文献中的每一个关键词相对于108个原子特征词的相似度值,然后乘以该关键词在摘要中出现的频次加1之后最大相似度值,即为加权相似度。

(4)构建文献空间向量矩阵:以681篇文献作为行,108个原子特征词作为列,以步骤(3)中求出的加权相似度值作为矩阵元素来构建文献空间向量矩阵 $\boldsymbol{R}$。该矩阵每一行即为一篇文献的一个空间向量,每一列(每一维)即为一个文献特征。通过相似

表1 原子特征词表

1	安全	19	传播	37	馆员	55	排架	73	数字	91	影响因子
2	版本	20	导航	38	集成	56	评估	74	搜索引擎	92	用户
3	版权	21	电子	39	计算机	57	评价	75	索引	93	语义
4	保存	22	调查	40	家谱	58	期刊	76	图书	94	元数据
5	被引	23	读者	41	价值	59	企业	77	图书馆	95	阅读
6	本体	24	敦煌	42	检索	60	情报	78	图像	96	整理
7	编目	25	分布式	43	建筑	61	全文数据库	79	推送	97	政府
8	标引	26	分词	44	借阅	62	人文	80	网络计量	98	知识
9	博客	27	分类	45	竞争	63	人性化	81	网页	99	智能
10	采访	28	佛经	46	口述	64	儒家	82	网站	100	中图法
11	采购	29	个人	47	类目	65	商务	83	文献	101	主题
12	参考	30	个性化	48	联合	66	社会	84	信息	102	著录
13	藏书	31	公共	49	联机	67	社区	85	虚拟	103	著作
14	查全	32	共享	50	联盟	68	史料	86	叙词表	104	专利
15	查新	33	古籍	51	链接	69	视频	87	学科	105	咨询
16	查询	34	关键词	52	论文	70	收录	88	学术	106	资料
17	成本	35	馆藏	53	目录	71	书目	89	引文	107	资源
18	出版	36	馆际	54	内容分析	72	数据	90	隐性知识	108	自动化

度加权计算的方法所取得的值与现有的方法相比较更加准确地反映出某个关键词在文献中的权重，而且所得的文献空间向量矩阵 R 中的数据的稀疏程度将大大降低，这样通过 FCM 聚类做出的结果将会更加精确、稳定。

(5)聚类实现：将文献空间向量矩阵 R 中的数据存入文本文档，作为 FCM 算法的数据源。运行 FCM 算法 50 以上，取出其中聚类目标函数值最小时的聚类结果作为最终结果输出。通过多次运行并对结果进行比较发现，聚类的结果比较稳定。

(6)学科交叉研究方法：通过 FCM 算法运行所得的学科分类数据同(2)中已知的学科分类数据相比较，得到一张学科交叉表。该表以原子特征词作为行，以两两学科相互之间是否交叉作为列，统计原子特征词是否在某两门学科之间出现以及如果出现，那么出现的频次为多大。这样就可以清晰地知道：哪些文献属于交叉学科，学科之间正在共同探讨哪些方面以及该方面的关注度如何。

5 聚类结果统计分析

5.1 结果统计方法

建立一张二维表，通过 FCM 聚类所得的文献分类结果与文献在现实中的学科分类相比，就可以很清楚地看到该文献是否为交叉学科，同时很清楚地看到该文献属于哪几门学科交叉及各学科之间的交叉点(关注点)。

表2 部分学科交叉表

原子特征词	Ⅰ-1	Ⅰ-2	Ⅰ-3	Ⅱ-1	Ⅱ-2	Ⅱ-3	Ⅲ-1	Ⅲ-2	Ⅲ-3
安全	0	3	0	0	2	0	5	0	0
版本	0	0	3	0	0	0	0	0	0
版权	0	0	0	0	2	0	4	0	0
保存	0	5	0	0	0	0	0	0	0
被引	0	2	4	0	0	0	0	0	0
本体	0	8	0	0	0	0	0	6	0

表2中每个字段中前一个数字表示文献通过FCM聚类方法聚成的结果，共分为3类，分别用Ⅰ、Ⅱ、Ⅲ表示；后一个数据表示文献的实际分类，其中“1”代表“图书馆学”；“2”代表“情报学”；“3”代表“文献学”。例如，“Ⅰ-1”前一个“Ⅰ”代表文献通过FCM聚类后分在第一类中，而后一个“1”表示文献现实中被归为图书馆学。从表2中我们可以看出：

(1)文献中哪些属于交叉学科范畴。当一部分文献通过原子特征词被聚为同一类时，说明这些文献研究或探讨的内容有相同或相似的方面，在这一类中的文献如果现实中属于不同学科时，那么说明学科之间有交叉的部分，其中一些文献属于交叉学科。从原子特征词中，我们可以进一步看出学科之间共同关注的课题。

(2)如果某个原子特征词分别在不同学科中出现，那么从中我们可以很明显地看出该原子特征词被哪些学科同时关注。例如原子特征词“编目”通过FCM聚类被分为同一类，但在现实中同时出现于“图书馆学”、“情报学”、“文献学”三门学科。从中我们可以看出，这三门学科正在同时研究“编目”这个方面。

(3)通过原子特征词的统计频次，我们可以进一步看出某个原子特征词的关注度。如果某个原子特征词在某些学科中同时多次出现，那么可以肯定地是这个原子特征词是不同学科研究的热点问题。例如：“检索”在图书馆学中出现的频次为7，在情报学中出现的频次为98，在文献学中出现的频次为11，而且是被聚为同一类中。从中可以看出，三门学科在“检索”方面属于交叉学科范畴，而且它在三门学科中的关注度很高。

5.2 统计分析

5.2.1 聚类结果的学科类别统计分析

表3 聚类结果的学科类别统计表

类别 \ 文献数	图书馆学		情报学		文献学		合计
Ⅰ	63	20.3%	154	49.7%	93	30%	310
Ⅱ	106	75.2%	32	22.7%	3	2.1%	141
Ⅲ	207	90%	21	9.1%	2	0.9%	230
合计	376		207		98		681

从表3统计的结果可以看出，聚类结果的第Ⅰ类主要是由情报学(占49.7%)和文献学(30%)组成，图书馆学占20.3%，因此第Ⅰ类中可以分析出情报学与文献学、情报学与图书馆学之间的交叉关系；第Ⅱ类中主要是由图书馆学(占75.2%)和情报学(占22.7%)组成，因此第Ⅱ类的结果可以分析出情报学与图书馆学之间的交叉关系；第Ⅲ类主要由图书馆学(占90%)组成，包含9.1%的情报学有可能成为研究的新的增长点。

5.2.2 学科研究热点分析

综合上述的统计情况，类别Ⅰ是三个学科的交叉，类别Ⅱ、Ⅲ主要是图书馆学为主，因此，按以下思路分别对三个类别进行统计分析来研究各个学科的研究热点：

(1)从第Ⅰ类中，按情报学中原子特征词频次降序排列后，排在前10位的数据如下：

原子特征词	图书馆学	情报学	文献学	总计
情报	0	115	0	115
检索	7	98	11	116
竞争	0	71	0	71
信息	19	65	0	84
数据	8	63	9	80
资源	27	52	20	99
数字	12	43	11	66
知识	14	43	0	57
参考	7	32	5	44
期刊	13	30	6	49

(2)从第Ⅱ类中，按文献学中原子特征词频次降序排列后，排在前10位的数据如下：

原子特征词	图书馆学	情报学	文献学	总计
文献	4	8	42	54
资源	27	52	20	99
编目	7	2	19	28
查新	0	7	15	22
分词	0	0	15	15
标引	0	3	13	16
著录	4	0	13	17
藏书	2	0	12	14
检索	7	98	11	116
数字	12	43	11	66

(3)从第Ⅲ类中,按图书馆学中原子特征词频次降序排列后,排在前10位的数据如下:

原子特征词	图书馆学	情报学	文献学	总计
图书	295	0	2	297
图书馆	292	0	2	294
数字	97	0	0	97
知识	53	16	0	69
资源	39	11	0	50
馆员	27	0	0	27
社区	25	0	0	25
评价	23	0	0	23
信息	17	0	0	17
社会	17	0	0	17

(4)研究热点汇总。将三张表的结果进行汇总得到每个学科研究的热点如下表:

研究热点 / 学科	研究热点
图书馆学	数字图书馆、知识管理、信息资源、信息评价、社区图书馆等
情报学	情报检索、竞争情报、知识管理、信息资源管理、信息数字化等
文献学	文献资源管理、文献检索、文献查新、数字文献、藏书等

5.2.3 学科交叉分析

从三张表中字体为黑斜体的数据统计出三个学科之间的交叉情况如下表:

图书馆学 VS 情报学	数字图书馆、知识管理、信息资源等
情报学 VS 文献学	文献检索、编目、文献、信息资源等
文献学 VS 图书馆学	编目、文献、信息资源、知识管理等
文献学 VS 图书馆学 VS 情报学	检索、数据、资源、参考、期刊、编目等

5.2.4 新的学科增长点分析

为了研究新的学科增长点,我们从第Ⅲ类中提取了情报学文献进行研究,由于该类别中绝大部分属于图书馆学,尽管其中只有21篇文献既属于图书馆学又属于情报学研究内容,因此该类别中这种学科交叉点有可能成为情报学新的增长点。具体数据见图2。

从上图可以看出,其中"数字图书馆"涉及较多,是2005年图书情报学研究的热点;而其中的"语义Web"、"本体"、"知识"等词的出现频次虽然较少,应该成为我们必须关注的信号,因为它们在若干年后可能成为图书情报学研究的新的增长点。为了验证我们的结论,我们从中知网的数字出版物超市《中国学术文献网络出版总库》学科学术热点,从中检索"本体"关键字的结果如下表:

4	ODL:一种基于本体的新型数字图书馆	数字图书馆/语义Web/本体/信息组织/信息集成/	3	数字图书馆 2; 语义Web 1; 本体 5; 信息集成 1;	2
9	基于数字化校园环境的图书馆门户建设	图书馆门户/数字化校园/信息资源管理/	3	图书馆门户 3; 数字化校园 3;	2
13	数字图书馆非图书馆	数字图书馆/图书馆/信息技术/数字资源建设/	3	数字图书馆 2; 图书馆 5; 信息技术 1; 数字资源建设 1;	2
46	高校图书馆网站评价指标体系研究	网站建设/评价指标/高校图书馆/	3	评价指标 3; 高校图书馆 2;	2
68	利用UML技术建立图书馆个性化推送系统模型	UML/个性化推送/图书馆/信息推送/	3	UML 1; 个性化 1; 图书馆 1;	2
1	论数字图书馆的知识构建	数字图书馆/知识构建/知识服务/知识网络/知识元/知识元	3	数字图书馆 4; 知识构建 3; 知识服务 1; 知识网络 1; 知识元 3;知识元链	2
107	走向互联的数字图书馆	Web Service/数字图书馆/SOAP协议/XML/网络互联/	3	Web Service 2; 数字图书馆 3; SOAP协议 0; XML0; 网络互联 2;	2
157	国内外图书馆数字参考咨询服务的比较研究	图书馆/数字参考服务/DRS/	3	图书馆 3; 数字参考服务 1; DRS	2
184	我国数字图书馆个性化信息服务探讨	数字图书馆/信息服务/个性化服务/	3	数字图书馆 2; 个性化信息服务 2;	2
224	论服务主导型数字图书馆体系结构与服务平台建设	服务主导型/数字图书馆/结构体系/服务平台/	3	服务主导型 3; 数字图书馆 2; 服务平台 2;	2
226	网络时代图书馆参考咨询服务新探	网络环境/参考咨询服务/图书馆/	3	网络环境 1; 参考咨询服务 2; 图书馆 2;	2
1	SFX技术在资源整合利用中的应用	图书馆/资源整合/SFX技术/动态链接/	3	图书馆 2; 资源整合 2;	2
277	论知识管理态势下的图书情报专业人才培养	知识管理/图书情报/专业人才培养/课程教学/	3	知识管理 3; 图书情报 2;	2
299	数字图书馆应用技术研究	数字图书馆/体系结构/元数据/检索技术/	3	数字图书馆 3; 元数据 1;	2
1	基于开放标准的数字图书馆检索接口	数字图书馆/分布式检索/J2EE/Web服务/	3	数字图书馆 4; 分布式检索 2; J2EE 1; Web服务 2;	2
368	数字图书馆的个性化推送服务	推送服务/个性化服务/数字图书馆/	3	推送服务 1; 个性化服务 1; 数字图书馆 2;	2
406	图书馆学情报学的持续发展——走向知识管理	图书馆学/情报学/知识管理/持续发展/	3	图书馆学 2; 情报学 1; 知识管理 2; 持续发展 1;	2
432	网络数字化图书馆资源建设探讨	数字图书馆/资源建设/数字化资源/	3	数字图书馆 2; 资源建设 2;	2
459	中美大学图书馆用户教育的发展研究	大学图书馆/用户教育/信息素质/网络信息资源/教育模式/	3	大学图书馆 4; 用户教育 3; 教育模式 1;	2
474	指纹识别技术在数字图书馆中的应用	图书馆/指纹识别技术/身份认证/	3	图书馆 4; 指纹识别技术 1; 身份认证 1;	2
533	浅谈图书馆网络信息资源建设	信息资源建设/图书馆/网络信息资源/	3	信息资源建设 2; 图书馆 1;	2

图2 第Ⅲ类中情报学文献数据

序号	热点主题	主要知识点	主题所属学科名称	热度值	文献数	相关国家课题数	研究人员数	研究机构数
5	语义网；知识服务；本体；	语义网；知识服务；本体；Web 服务；xml；知识管理；万维网；高校图书馆；rdf；元数据；网络检索；图书馆；图书馆员；本体映射；智能检索；图书馆服务；知识地图；知识服务系统；数字图书馆建设；语义 Web	图书情报与数字图书馆；计算机软件及计算机应用	10149	178	25	273	127
6	语义网；本体；信息组织；	语义网；本体；Web 服务；数字图书馆；万维网；rdf；xml；文献标题；搜索引擎；网络检索；元数据；知识服务；智能检索；圆周率；本体论；知识表示；信息检索系统；ibase 数据库；ontology；本体语言	互联网技术；计算机软件及计算机应用	7959	146	28	226	104

从表中可以检验我们对 2005 年数据的挖掘和预测的结果在 2009 年得到验证。

6 结 论

科技文献反映了学科研究的最新成果和研究内容，通过文献聚类可以发现学科交叉和学科热点等隐藏在数据背后的知识。学科交叉研究对于学科及研究者都具有十分重要的意义，本文提出了一种通过文献聚类的方法来进行学科交叉研究，为学科交叉的研究提供了一种新的思路。通过数据预处理、建模、实验仿真、结果的统计分析和结果检验等过程，说明本文提出的方法是科学的、可靠的。本文主要的创新点：①通过文献数据的聚类分析来研究学科交叉。②在设计的 VSM 模型中，提出了一种新的相似度加权模型，从而使得文献聚类所得的结果更加合理和准确。③在聚类结果的分析中，设计了学科交叉表，可以一目了然地看出哪些学科在哪些方面交叉，同时还可以看出哪些课题是学科之间的热点及新的增长点。这样研究人员、学者就可以在第一时间很清楚地了解到交叉学科的热点及增长点，不仅节省花在查找方面的时间与精力，更为重要的是，可以时刻掌握先机，抢占科学前沿。

参 考 文 献

[1] 路甬祥.学科交叉与交叉科学的意义[J].中国科学院院刊，2005(01):58-60.

[2] 金薇吟.学科交叉理论与高校交叉学科建设研究[D].苏州大学，2005.

[3] 赵晓春.跨学科研究与科研创新能力建设[D].中国科学技术大学，2007.

[4] 李春景，刘仲林.现代科学发展学科叉模式探析——一种学科交叉模式的分析框架[J].科学学研究，2004(03):244-248.

[5] 彭小平.对学科交叉的探讨与研究[J].科技情报开发与经济，2008(04):158-160.

[6] 王庚华，邱岩，谢寅波.大学学科交叉的路径[J].中国冶金教育，2008(04):5-8.

[7] 顾浩.论学科交叉路径及趋势[J].上海金融学院学报，2006(06):67-69,73.

[8] 李喜先.论交叉学科[J].科学学研究，2001,19(1)：22-27.

[9] 李钢，汤仲胜.我国交叉学科发展现状与趋势研究[J].科学学与科学技术管理，2000,21(11)：45-49.

[10] 曹付元，梁吉业，姜广.基于邻域模型的 K-means 初始聚类中心选择算法[J].计算机科学，2008,35(11)：181-184.

[11] 索红光，王玉伟.一种用于文本聚类的改进 k-means 算法[J].山东大学学报(理学版)，2008，43(1)：60-64.

[12] 宋江春，沈钧毅.一个基于双向近邻技术的多层文档聚类算法[J].情报学报，2006，25(4):488-492.

[13] 吴景岚，朱文兴.基于 K 中心点的文档聚类算法[J].兰州大学学报(自然科学版)，2005，41(5):88-91.

[14] 白曦，吕晓枫，孙吉贵.融合模拟退火的遗传算法在文档聚类中的应用[J].计算机工程与应用，2006，42(23):144-148.

[15] 张云，冯博琴，麻首强，等.蚁群——遗传融合的文本聚类算法[J].西安交通大学学报，2007，41(10)：1146-1150.

[16] 林春燕，朱东华.科学文献的模糊聚类算法[J].计算机应用，2004,24(11):66-70.

[17] 孟海涛，陈笑蓉.基于模糊相似度的科技文献软聚类算法[J].贵州大学学报(自然科学版)，2007，24(2)：175-178.

[18] 魏建香，苏新宁.基于关键字和摘要相关度的文献聚类研究[J].情报学报，2008，28(2)：220-224.

(责任编辑 芮国章)