● 魏建香[1]，孙越泓[2]，苏新宁[3]
(1. 南京人口管理干部学院 信息科学系，江苏 南京 210042；2. 南京师范大学 数学科学学院，江苏 南京 210097；3. 南京大学 信息管理学院，江苏 南京 210093)

学科交叉知识挖掘模型研究*

摘 要：为揭示学科之间的交叉知识，提出一种基于文本挖掘的学科交叉知识发现模型。构建了学科交叉文献发现模型与学科交叉知识发掘模型，从学科交叉点和新的增长点两个方面来揭示学科之间的交叉关系，并通过近10年来情报学与计算机科学两个学科文献数据进行了实例验证。

关键词：学科交叉；知识挖掘；文本挖掘；知识发现；模型

Abstract: In order to disclose the interdisciplinary knowledge among disciplines, this paper proposes an interdisciplinary knowledge discovery model based on text mining. The paper constructs the interdisciplinary literature discovery model and the interdisciplinary knowledge excavation model, discloses the cross relationships among disciplines from the disciplinary intersection point and new growth point, and uses the literature data of information science and computer science in recent 10 years as examples to validate the results.

Keywords: interdiscipline; knowledge mining; text mining; knowledge discovery; model

从信息时代走向知识时代是信息社会发展的必然趋势。通过对国内文献数据库的利用情况和学科研究现状分析，有两个方面的需求亟待解决：①虽然国内中文科技文献数据库系统已经日趋完善并积累了海量的文献数据，但各文献数据库系统对数据的开发和利用却非常薄弱，主要提供基本的信息检索和简单的统计功能，不能上升到知识发现的层面。因此经常用“淹没于信息，饥渴于知识”来形容这种无奈。②随着科学技术的进步，学科之间的界线逐渐被打破，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。学科之间交叉融合为学科发展提供了动力，跨学科研究成为学科研究的前沿和热点。许多学者都存在这样的困惑：本学科与其他学科的交叉点在哪，学科增长点在哪，如何从海量文献数据中发现这样的知识？文本挖掘是数据挖掘的一个重要分支，是指从大规模文本库中抽取隐含、以前未知、潜在有用模式的知识发现过程。文本挖掘可以有效地发现超越文本本身的有用知识，如学科之间研究的交叉知识，它为以上需求提供了技术上的可行性。本文利用文本挖掘技术，构建用于发现学科之间交叉知识的挖掘模型，为揭示学科交叉提供一种新思路。

1 研究现状

对于学科交叉的概念，许多学者都提出了自己的看法和见解：中国科学院院士路甬祥认为，学科交叉是“学科际”或“跨学科”研究活动，其结果导致的知识体系构成了交叉科学[1]；刘仲林在《现代交叉科学》一书中指出，所谓学科交叉，指跨出已有学科的边界，实现学科间的合作。换句话说，凡打破已有学科壁垒，把不同学科理论、方法或思维有机地融为一体的研究活动，就是学科交叉[2]。

国外也将学科交叉称为跨学科。国外相关研究起始于20世纪60年代后期，Koester于1968年编著了第一次国际跨学科研讨会会议论文集《超越还原论：阿尔巴赫问题论丛》，标志着跨学科研究的开始。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。1980年，国际跨学科协会正式成立，并在美、英、德、法等国召开跨学科研究学术会议，标志着跨学科研究在世界范围内兴起。1990年美国跨学科学专家克莱茵(J. T. Klein)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[3]。

我国学科交叉研究萌生于20世纪50年代，到80年

* 本文为国家社会科学基金青年自选项目（项目编号：09CTQ022），教育部人文社会科学重点研究基地2008年度重大项目（项目编号：08JJD870225）和江苏省“六大人才高峰”第六批资助项目（项目编号：09-E-016）的研究成果之一。

代进入全面展开阶段。主要著作包括：徐纪敏编著的《科学的边缘》，李光与任定成主编的《交叉科学导论》，刘仲林主编的《跨学科学导论》、《跨学科教育论》、《现代交叉科学》。中国社会研究院的王兴成将我国跨学科研究分为4个阶段：孕育阶段（20世纪20—40年代）、起步阶段（20世纪50—60年代）、发展阶段（20世纪70—80年代）、提高阶段（20世纪80—90年代），概括了我国交叉学科研究的发展历程及启示[4]。国内利用数据进行学科交叉研究的文献有：邱均平等通过对2001—2003年的CSSCI文献计量分析研究，得出图书馆、情报与档案管理同新闻与传播学、经济学、教育学和法学等学科有比较多的交叉研究[5]。杨建林等利用文献之间的引文关系来研究情报学和其他学科的交叉信息[6]。于江等以2000—2005年度有关部门管理科学相关领域项目申请书的分析数据为依据，对我国基础科研领域发展状况作了分析，重点对研究热点的识别、学科的交叉、学科的演化趋势作了分析研究[7]。以上研究主要采用文献计量分析方法，缺乏基于文本挖掘的方法，因此本研究具有一定的价值。

2 学科交叉文献发现模型

要挖掘两个学科之间的交叉关系，首要的问题是发现哪些文献属于跨学科研究文献。在现有的文献数据集中，从期刊的学科特性能够比较直观地了解每一篇文献所属的学科，从文献的中图分类号，也能大致了解哪些文献属于两个学科的交叉文献。如《情报学报》期刊中中图分类号为“TP3”的文献为计算机学科文献。但由于中图分类号是由作者或编辑人为给定的，主观性强，因此采用这种方式来进行学科交叉文献的识别是有缺陷的。为了克服该方法具有人为主观意志的缺陷，需要找到一种客观、基于数据本质特征的交叉学科文献识别方法。聚类算法可以将文献进行分类，但存在以下问题：①由于目前聚类结果都是类别无交叉，因此从聚类结果找到符合学科交叉特征的文献存在困难。②由于文献量大，文献聚类将在一个超高维的空间中进行，因此效率低下。问题的根本在于：一方面要将数据分类；另一方面要将不易区分类别的数据找到。对于数据分类不需要采用所有关键词聚类，只需要构造3个基本特征，即学科I特征、学科II特征和学科交叉特征，就可以将所有文献区分；由于学科交叉文献本质上是属于模糊问题，因此可以借助模糊C均值算法（FCM）中的隶属度函数来发现难以区分类别的文献。隶属度函数是FCM算法中非常有用的一个度量，表示一个对象属于某个类别的概率，可以依据该度量值构造模糊函数，通过模糊函数的阈值设置来控制一篇文献是否属于学科交叉文献。学科交叉文献发现流程见图1。

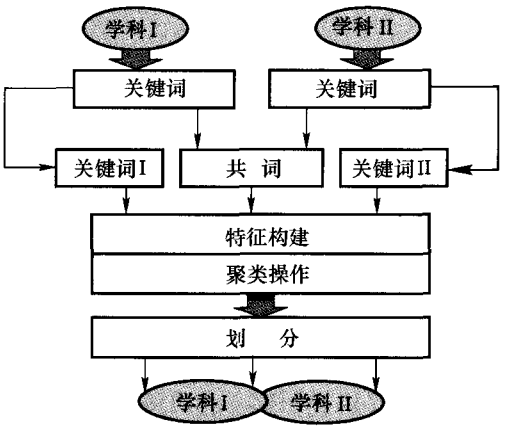


图1 基于聚类的学科交叉文献发现模型

定义1：假设文献集A共包含两个学科，分别记为学科Ⅰ与学科Ⅱ。依据期刊的学科特性将A划分成两个子集A_1与A_2，如果$\exists\varphi$，使得$A_1 \cap A_2 \xrightarrow{\varphi(U_1, U_2)} C$，则称$C$为两个学科的交叉文献，其中$\varphi$是基于模糊隶属度函数$U$的一个模糊划分。

定义2：假设K_1为A_1的所有关键词集合，K_2为A_2的所有关键词集合，$K_3 = K_1 \cap K_2$，$K_{1\_3} = K_1 / K_3$，$K_{2\_3} = K_2 / K_3$，则称K_1为学科I特征词，K_2为学科II特征词，K_3为学科交叉特征词。显然，$K_{1_3} \cap K_3 = \Phi$且$K_{2_3} \cap K_3 = \Phi$。

定义3：评价函数G为φ划分的评价标准，G定义如下：

$$G = F_1 + F_2 = \frac{2P_1R_1}{(P_1 + R_1)} + \frac{2P_2R_2}{(P_2 + R_2)} \qquad (1)$$

其中P_i，R_i，F_i为$i(i=1, 2)$类中的查全率、查准率和F指标。G为两个类别的F值之和，φ划分的原则是既要保证一定数量的划分到交叉类别中，但又要确保G值较高。

在上述定义的基础上，设计出交叉学科文献的发现算法：①依据期刊的学科特征标志所有文献的类别号。②提取每个类别文献的关键词，构建学科特征关键词集合K_{1_3}，K_{2_3}和K_3。③分别计算每一个特征词集合中的关键词在每一篇文献的标题、中文关键词和中文摘要出现的频次之和并求出频率。④构建VSM矩阵，以文献为行，3个特征为列构建矩阵，矩阵元素为频率（文献对特征的贡献度）。⑤规格化数据，由于每个元素代表了它在每一列中出现的频率，而3个特征空间的维度不同，对数据规格化以消除量纲的偏差。⑥将规格化后的VSM矩阵采用FCM算法进行聚类。⑦对聚类后的结果，调整隶属度函数，当评价函数最优时，输出结果。

3 学科交叉知识发掘模型

在找到学科交叉文献的基础上，如何确定两个学科共同的研究方向及新的研究点，是本文要重点解决的问题。通过聚类分析可以使相似性更高的文献聚为一类，在共词聚类中，具有相同关键词的文献聚为一类，可以通过基于关键词的类别特征提取来描述该类别，该类别即为学科交叉研究的共同点。由于新的研究方向通常由突现词的文献来体现，因此对于学科交叉中可能的新方向可以利用突现词文献聚类分析获得。为此，从共同的研究方向与新的研究热点两个方面来构建学科交叉知识发现模型。

3.1 基于共词聚类的学科交叉点挖掘模型

文本挖掘的最终目标是知识发现和利用。在学科交叉研究中，通过共词聚类分析的目标是挖掘学科交叉的研究方向，即学科交叉点。交叉点的发现需要对交叉文献进行深度挖掘，以深入地了解学科间共同的研究方向。图 2 给出了基于共词聚类的学科交叉点挖掘模型。

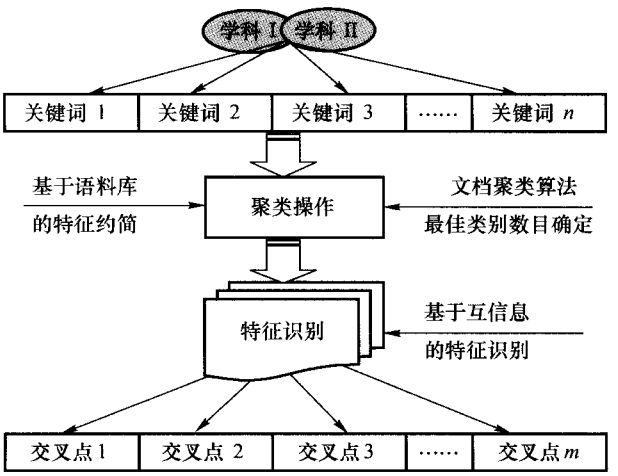


图 2 基于共词聚类的学科交叉点挖掘模型

模型说明如下：①从学科交叉文献中抽取所有文献的关键词，并将低频关键词剔除。②由于特征词数量极大，在特征损失较小的前提下，通过降维技术将关键词降到可以接受的数量。③将经过降维后得到的关键词作为聚类操作的特征词，采用性能较好的聚类算法（如 FCM 算法）进行聚类。由于聚类的类别数目确定了交叉点的数量，因此对类别数目事先采用最佳类别评估方法（如基于混合 F 统计量[8]）进行确定。④学科交叉点的识别方法可以用基于互信息的特征提取算法方法对每一个类别进行特征描述，该描述即为学科交叉共同的研究方向。

3.2 基于互信息的研究方向描述

在对学科交叉文献进行聚类操作后，需要提取类别的本质并对类别进行有效描述。互信息是表示两个变量之间关联程度有用的度量，因此可以通过关键词与类别的互信息来进行类别的特征提取。在对学科交叉文献聚类后，具有相同研究方向的文献聚为一类，这些文献在特征词上具有共性，提取这个共性词汇就是对研究方向的描述。例如：如果一个类别中大部分文献的关键词是“搜索引擎”或相关词汇，而其他类别的文献中很少包含“搜索引擎”，则可以认定该类别的研究方向名称为“搜索引擎”。也就是说，与其他类别相比，“搜索引擎”与该类别的互信息最大。由于与同一个类别互信息最大的关键词不只一个，因此类别特征有可能由一组词汇组成。基于互信息的类别特征描述算法如下：

1）统计聚类结果中总的文献数 n 与各类别文献数 n_i，$i=1,2,\cdots,k$（k 为聚类数目）。

2）打开学科研究方向语料库表（该语料库包含了研究方向特征词，其中一个研究方向由多个关键词组成，而第一关键词为主关键词），扫描所有记录。

3）计数器 x——统计主关键词出现的总频次，类别统计 $x_i(i=1,2,\cdots,k)$ ——统计主关键词在各个类别中出现的频次，置所有计数器初值为 0。

4）置第 1 字段为主关键词，记为 t_k，其他字段为次关键词，记为 $p_j(j=1,2,\cdots,m,m$ 为次关键词个数）。

5）打开聚类结果表，并将 t_k 及所有的 p_j 与每一条文献记录的每一个关键词进行比较，如果相同，则 $x=x+1$，并将将第 i 类的类别计数器 x_i 加 1。

6）扫描聚类结果表中所有记录，则语料库中的第 1 条记录的主关键词与所有类别的互信息（MI）可以用以下公式计算得到：

$$\mathrm{MI}(t_k,c_i)=x_i\cdot\log\frac{P(t_k,c_i)}{P(t_k)P(c_i)}=x_i\cdot\log\frac{(x_i/n_i)}{(x/n)(n_i/n)},\quad i=1,2,\cdots,k \tag{2}$$

公式（2）表示主关键词的互信息值乘以该类别中与主关键词相关的文献篇数。主要有两方面的考虑：①为了突出类别中文献多且互信息高的关键词。②由于互信息既考虑了关键词在目标类别中出现的概率，又兼顾了在其他类别中出现的概率，当某个关键词只在一个类别中出现时，该关键词相对于这个类别的互信息会异常高，但有时包含该关键词的文献在该类别中并不占主导地位。用相关文献篇数加权可以有效地避免类似情况发生。

7）重复 3）~6），求出所有主关键词与所有类别之间的互信息值。

8）分类别统计最大互信息值及对应的主关键词。

3.3 学科交叉突现词检测方法

新概念的出现有可能代表了一个新的研究方向。目前，已经有许多学者在知识发现中开始注意到新术语的出现，例如，陈超美博士在 CiteSpace 软件中利用突现词

(Burst Terms) 探测，来分析学科的前沿领域和发展趋势[9]。突现词对于学科交叉研究意义重大：一方面，学科交叉是多学科的交融点，通过多个学科知识火花的碰撞，最有可能涌现新的突现词。另外，突现词也会给各个学科的发展带来新的发展方向和机遇，从而促进各个学科间的进一步交融。由于突现词有可能成为学科新的增长点，在学科交叉研究中，通过学科突现词的挖掘给研究者在选择研究方向时提供帮助。突现词的检测方法设计如下：①确定一个若干年的研究周期和多个学科，并选择一定量的有代表性的学科期刊文献。②将文献按年份分组，并求出各年份的关键词集合。③对每个学科，从第二年开始将每一年的关键词集合与前面所有年份的关键词集合进行比较，如果有关键词未在前面所有年份中出现，则该关键词为该年份的新术语。④通过第 2 节中学科交叉文献发掘模型，提取学科交叉特征词集合，如果新术语属于学科交叉特征词集合，则保留；否则剔除。⑤设置一个阈值，如果保留的新术语在后续年份中出现的频次达到该阈值，则确定该术语为突现词。设置阈值的做法可以避免一些冷词或虚泛词的干扰。

4 实证研究

为了验证学科交叉知识挖掘模型的可行性，选择了情报学与计算机科学两个学科近 10 年的 12 944 条核心期刊文献题录数据进行实例研究。

4.1 学科交叉特征词选择

样本文献中共包含 52 842 个关键词，平均为 4.08/篇，互异关键词为 24 621 个。定义情报学相关期刊文献为类别 I，共 7 139 篇，计算机相关期刊文献为类别 II，共 5 805 篇，分别存入两张表中，并求两类学科特征词集合 K_1 和 K_2。从 K_1 和 K_2 中选择共同的特征词，构成学科交叉特征词。频次排名前 10 的关键词见表 1。

表 1 学科交叉文献关键词频次（TOP 10）

序号	关键词	频次	序号	关键词	频次
1	信息检索	465	6	数据库	163
2	搜索引擎	321	7	网络安全	109
3	数据挖掘	262	8	元数据	103
4	本体	234	9	Internet	101
5	XML	193	10	网格	100

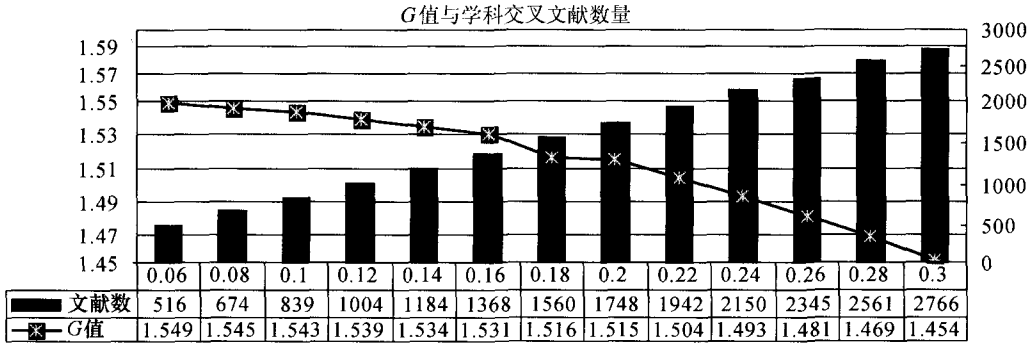
4.2 学科交叉文献选择策略

在确定 3 个特征词集合后，根据交叉学科文献的发现算法步骤构建文献的 VSM 矩阵，并采用 FCM 算法进行聚类（分两类），得到每一篇文献属于两个类的隶属度。为了求出学科交叉文献，需要设计一个模糊函数。

定义 4：对每一篇文献 x，属于类别 C_1 和 C_2 的隶属度分别记为 u_1，u_2，$u_1+u_2=1$；$\Delta u=|u_1-u_2|$；则模糊函数 φ 定义如下：

$$\varphi(U_1,U_2)=\begin{cases} x\in C & \Delta u\leqslant\alpha \\ x\in C_1 & \Delta u>\dfrac{1+\alpha}{2} \\ x\in C_2 & \Delta u<\dfrac{1+\alpha}{2} \end{cases} \tag{3}$$

α 为阈值，C 为交叉类别。根据上述定义，对划分结果根据阈值 α 从 0.06～0.3 进行变化。阈值 α 是代表了每篇文献属于两个类别的隶属度之差。选取原则是既要保证一定的交叉文献数量，但又要考虑到指标 G 保持一定的水平。图 3 给出了 G 值随 α 的变化情况，从图 3 可以看出，当 α 从 0.06 变化到 0.16 过程中，G 值的变化相对平稳；当 α 从 0.18 开始，G 值出现下降，即两个类别的 F_1，F_2 指标之和出现大的下降。根据以上分析，α 选择 0.16 能较好地满足要求。取 $\alpha=0.16$ 时，“学科交叉类别”列中的文献作为学科交叉文献，其中包含 657 篇情报学文献和 711 篇计算机文献，共计 1 368 篇文献。



	0.06	0.08	0.1	0.12	0.14	0.16	0.18	0.2	0.22	0.24	0.26	0.28	0.3
文献数	516	674	839	1004	1184	1368	1560	1748	1942	2150	2345	2561	2766
G值	1.549	1.545	1.543	1.539	1.534	1.531	1.516	1.515	1.504	1.493	1.481	1.469	1.454

图 3 评价函数 G 与学科交叉文献量变化图

4.3 类别特征描述

对 1 368 篇两个学科交叉文献进行聚类操作，并通过第 3.2 节中公式（2）对类别进行描述，结果见表 2。

从表 2 可以看出，情报学与计算机科学在 10 年的研究中，共同的研究方向为“数据挖掘”、“搜索引擎”、“软件工程与图像处理”、“信息检索”、“信息安全”与“本体”6 个方向，其中“本体”是近几年正在兴起的交叉点。在对 10 年数据逐年增量聚类的基础上，交叉研究方向形成过程见图 4。

表2 交叉研究方向描述

类别	互信息（TOP 5）	类别描述
1	信息安全：263.877；网络：84.4041；软件工程：72.6533；服务：60.5485；信息管理：51.4492	信息安全
2	信息检索：634.035；语义：193.255；本体：170.887；多媒体：78.2462；语义网：67.7407	信息检索
3	软件工程：69.5612；图像：58.0961；信息安全：30.7636；算法：18.89；AGENT：16.2706	软件工程与图像处理
4	数据挖掘：453.076；数据库：86.9509；关联规则：71.0434；聚类：70.2904；XML：40.9205	数据挖掘
5	本体：164.294；图像：109.375；XML：61.2113；网络：54.5972；算法：25.047	本体
6	搜索引擎：604.616；信息检索：90.3371；网络：81.2729；元数据：69.198；用户：34.8539	搜索引擎

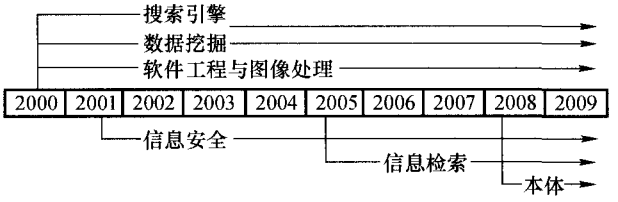


图4 学科交叉研究方向形成过程

4.4 突现词聚类分析

根据第3.3节中突现词的检测方法，在1 368篇交叉学科文献中共计检测出实发词79个，其中频次大于10的突现词信息见表3。对包含上述突现词的146篇文献采用FCM算法进行聚类，结果见表4。

表3 高频突现词相关信息

序号	年份	突现词	频次	篇名	刊名
1	2002	相关反馈	17	Internet个性化智能信息检索的分析与研究	情报学报
2	2002	Google	13	Google搜索引擎PageRank技术的优化	情报科学
3	2003	RDF	16	基于XML的3个常用元数据描述工具的评价与比较	情报科学
4	2002	个性化	20	Internet个性化智能信息检索的分析与研究	情报学报
5	2001	图像检索	27	基于内容的图像检索系统研究	现代图书情报技术
6	2003	领域本体	16	基于领域知识重用的虚拟领域本体构造	软件学报
7	2003	Web服务	20	WebMark：一个Web服务器性能测试工具	软件学报
8	2000	语义网	22	基于语义网络的概念检索研究与实现	情报学报
9	2004	语义Web	12	本体论研究综述	计算机研究与发展
10	2003	语义检索	15	基于多层次概念语义网络结构的中文医学信息语义标引体系和语义检索模型研究	情报学报

表4 高频突现词文献聚类结果

类别	类别特征描述	情报学	计算机	文献总数
1	Web服务	9	11	20
2	图像检索	21	9	30
3	领域本体	22	1	23
4	个性化	14	3	17
5	Google/RDF	33	3	36
6	语义网	17	3	20

从表4可以看出，聚类后形成了6个研究方向，每个研究方向的文献数在20篇左右。这6个研究方向中，情报学在“图像检索”、“领域本体”、“个性化服务”、“Google”、“语义网”的研究中发文量占有较大比例，有可能成为情报学与两个学科交叉新的增长点。

5 结束语

本文提出了一种基于文本挖掘的学科交叉研究方法，构建了学科交叉文献发现模型与学科交叉知识挖掘模型，从学科交叉点和新的增长点两个方面来揭示学科之间的交叉关系，并通过近10年来情报学与计算机科学学科文献数据进行了实例验证，为学科交叉知识发现提供了一个新的研究思路。下一步的工作将在本文提出的模型基础上，研究学科交叉知识可视化模型。□

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（1）：58-60.
[2] 刘仲林．现代交叉科学［M］．杭州：浙江教育出版社，1998.
[3] 金薇吟．学科交叉理论与高校交叉学科建设研究［D］．苏州：苏州大学，2005.
[4] 王兴成．跨学科研究在中国：历程和启示［J］．科学学研究，1995，13（2）：1-7.
[5] 邱均平，马瑞敏．基于CSSCI的图书馆、情报与档案管理一级学科文献计量评价研究［J］．中国图书馆学报，2006（1）：24-29.
[6] 杨建林，孙明军．利用引文索引数据挖掘学科交叉信息［J］．情报学报，2004，23（6）：672-676.
[7] 于江，党延忠．用信息可视化方法分析科研领域发展状况［J］．科学学与科学技术管理，2009（6）：10-14.
[8] 孙才志，王敬东，潘俊．模糊聚类分析最佳聚类数的确定方法研究［J］．模糊系统与数学，2001，15（1）：89-92.
[9] CHEN C. CiteSpace II：detecting and visualizing emerging trends and transient patterns in scientific literature［J］. Journal of the American Society for Information Science and Technology，2006，57（3），359-377.

作者简介：魏建香，男，1971年生，副教授，博士。研究方向：智能信息处理与文本挖掘。
孙越泓，女，1972年生，副教授，博士。研究方向：人工智能，图像处理研究。
苏新宁，男，1955年生，教授，博士生导师。研究方向：信息处理与检索，知识管理，引文分析等。

收稿日期：2011－09－16