

● 任红娟<sup>1,2,3</sup>, 张志强<sup>1,2</sup>

(1. 中国科学院 国家科学图书馆兰州分馆, 甘肃 兰州, 730000; 2. 中国科学院 资源环境信息中心, 甘肃 兰州 730000 3. 中国科学院 研究生院, 北京 100049)

## 基于文献内容和链接融合的知识结构划分方法研究进展

**摘要:** 知识结构划分是确定研究主题、发现新兴研究领域和识别重点研究内容的一个重要方法。本文把从文献内容特征和文献链接信息融合的新知识结构划分方法的研究进行了梳理, 从数据库扩展的原始级信息融合、文本挖掘和文献计量方法结合、词汇引用图和词参考文献共现 4 个层面对当前的知识结构融合方法进行了综述, 分析了它们的主要研究内容和方法, 并对目前知识结构划分的聚类方法进行了综合研究, 以期从更加新的视角对知识结构划分研究提供新的研究思路。

**关键词:** 文献内容; 信息融合; 知识结构; 研究进展

**Abstract:** Division of intellectual structure is an important method to determine research topics, find emerging research fields and identify key research content. This paper reviews the researches on the new intellectual structure division method which fuses the features of document content with the information of document link, summarizes the present intellectual structure fusion method in terms of the elementary information fusion to expand the database, the combination of text mining and bibliometrics, the term citation graph and the word reference co-occurrence, and analyzes their main research content and method. The paper also makes a comprehensive study of the present clustering method to divide intellectual structure in an attempt to provide a new research thinking for intellectual structure division from a newer angle of view.

**Keywords:** document content; information fusion; knowledge structure; research progress

从科学文献当中发掘科学的历史沿革、发展趋势、主要研究内容、突出的贡献者、知识结构等一直是科学计量学和文献计量学的重要研究内容。而科学文献有许多不同特征项, 包括标题、摘要、关键词、作者、正文、参考文献等, 这些不同的特征项包含着不同程度的文献的相关信息, 既有重合, 又有差异。从现有的研究方法来看, 关于知识结构的研究主要是基于各种共被引分析方法或者共词、引用等单一的方法为主。然而从文献作为科学交流的成果角度而言, 它并不是单一的特征就能够很好地表征的。例如, 两篇文献在用词上非常类似, 都采用“Opinion Analysis”, 其中一篇是作者对某个问题的主观的基于自身知识和经验的分析, 而另一篇则是研究网络当中的观点挖掘的, 两篇研究内容完全不同, 如果从其参考文献入手就会发现它们之间的相似性非常低。同样如果两篇文献通过共被引和耦合关系表明了相似性, 而它们可能是分属不同领域的方法被其他的文献所引用, 而从这两篇文献的用词上就会发现它们其实并不相似。

相对来说, 把文献的内容和链接(即引用)关系结合在一起进行的研究比较少, 而这种从整体上更加全面地

分析文献的特征, 并根据不同的特征信息结合在一起来进行知识结构划分更加准确已经得到了初步的确认<sup>[1]</sup>。近几年来关于信息融合的知识结构划分方法研究逐步受到了研究者的关注, 虽然相对大量的知识结构划分方法研究它们的比例还非常的小, 但是作为对传统的知识结构划分方法改进的一种多视角的研究方法, 值得我们进一步去探索, 以期能够为文献计量学的发展带来新的生机。

### 1 相关概念

#### 1.1 文献的内容和链接特征项

文献的内容特征项主要分为狭义和广义两种。狭义的内容特征项包括描述文献内容相关的词, 包括文献标题、摘要、关键词(作者关键词和增补关键词)、正文。文献的作者文献计量研究当中并不是作为个人主体特征来研究的, 两个作者的关系相近表明他们发表的作品的的内容相似, 这里的作者关系是指带作者作品之间的关系<sup>[2]</sup>。从这种意义上来说, 作者也是文献的内容特征描述项, 因此广义的内容特征项也包含作者, 甚至还包含期刊和机构在内。本文的内容特征限定为狭义的内容特征项。

文献除了包含一些描述内容的词汇以外，还包含参考文献，这也是科学文本区别于其他自由文本的一个重要特征。参考文献作为记录文献关系的特征项，在一定程度上可以反映文献研究内容的相似性，但它更多地可以反映文献与外部其他文献之间的链接关系，是存在于文献内部的外部链接“天线”。它可以从不同于词间关系的角度来挖掘文献以及文献其他特征项之间的关系。

### 1.2 融合的概念

数据融合分为原始级融合、特征级融合和决策级融合。其中原始级融合是最低层次的融合，是在采集到的传感器的原始信息层次上未经处理或者只做很小的处理进行融合，特征级融合是指从原始信息中抽取特征信息进行综合分析和处理，而决策级融合是将多个传感器的识别结果进行融合。融合的目的是为了能够更加全面和准确地分析问题和作出更加正确的决策。笔者提出的融合概念类似于数据融合概念的内涵，对应于文献不同特征项的数据融合、特征融合以及结果融合。这两种融合在形式上看起来非常类似，但是并不是等同的关系，它们有着本质上的区别。信息或者数据融合主要研究多源数据的融合，强调数据来源的多样性和数据的外源性，而基于文献特征项的信息融合主要是从文献内部的不同部分来进行融合的，强调数据分析对象的多样性和数据的内源性。

### 1.3 知识结构

知识结构的概念至今还没有明确的界定。1981年，Small和Griffith提出了作者共被引分析方法是研究知识结构的一种重要方法。从作者的研究来看，知识结构是专业或者学科的子领域或者子结构，是基于一定的文献特征项得到的数据集反映的学科或者专业的主题结构关系。从主题结构的研究来看，主要是从文献的词特征得到的主题分类或者学科结构，而知识结构的划分采用词关系和引用关系的都有。因此，作者认为知识结构包含主题结构，比主题结构的外延更广，可以包含宏观的国家或者整个大学科的主题结构、中观的一个学科的主题结构以及微观的一个专业或者一个研究主题的结构划分。基于作者关系得到的知识结构和无形学院有一定的相似关系，但是无形学院更加强调学者之间的社会维度的联系，有时在基于文献得到的知识结构关系中，这种无形学院关系是无法反映的。知识结构的内涵界定为：在确定的数据集中根据对象之间的关系得到的对象之间的亲疏关系的划分。

## 2 不同层面的信息融合知识结构划分方法研究

### 2.1 不同层面的领域分析数据的融合研究

2.1.1 基于多个数据源的信息融合方法研究 Synnes ved 的博士论文采用了信息融合的思想和方法<sup>[3]</sup>，利用概率

记录链接方法和信息融合方法来进行领域知识结构研究的数据准备，把WOS和Medline中有关生物医学的引用信息整合在一起，以期能够改进和丰富生物医学引用数据的可视化表示。研究表明，这种融合了多个数据库引用信息的信息融合方法能够提高数据的质量，增加可视化中的突发词以及改变关键词的等级排序，减少单一引用数据库造成的偏见，形成更加丰富的信息空间。这种信息融合是一种多源数据的结合，类似数据融合中的原始级和特征级融合方法的结合，是采用扩大数据集并把两种数据特征信息合并的方法，以期提高知识领域分析的全面性和准确性。

从文献计量的知识结构划分方法来看，利用一定的记录链接方法来进行多数据源融合的研究较少，多数研究还是采用从不同的数据库直接获取多个数据，并把它们作为一个整体数据集进行分析，例如Eam为了研究决策支持系统1971—1990年20年的知识结构，从3个来源获取了分析的数据<sup>[4]</sup>。因为大多数知识结构研究都是基于共被引数据来进行分析的，所以大多数研究都是在ISI数据库中直接通过关键词或者期刊引用报告JIR来获取领域相关数据的。也有一小部分作者从网络数据库，例如CiteSeer、Google Scholar或者通过一些专业的数据库例如Medline等来进行数据的选取。除了宏观上数据库的选择采用信息融合的方法以外，从微观上对于领域数据集的构建在一个数据库内可以基于不同的关键词、选择多种期刊以及利用引文和参考文献来进行数据集的扩大。Zit和Bassecouard利用关键词搜索的结果作为种子数据集<sup>[5]</sup>，然后再利用引用和被引的阈值来扩大数据集，并在扩大的过程中利用阈值来进行收敛，证明了这种融合方法的有效性。

2.1.2 基于文本挖掘和文献计量方法结合的信息融合方法 随着数字化和网络化的发展，科学文献信息以指数量级在不断增长，面对如此庞大的信息宝库，如何从中快速而准确地识别重要和关键的信息成为科学创新的关键一环。而数据挖掘、文本挖掘技术的出现和发展，使得快速和有效地处理大量的信息不再是梦想。传统的文献计量学方法由于受到处理方法的限制，主要把内容的分析限定在标题、摘要和关键词以及标引词这些对象上，并利用词频统计的方法选取高频词汇作为分析对象来进行领域知识结构研究。这种方法虽然在实践中已经被很多学者证明了有效性，但是这种方法本身并不是没有问题。例如阈值选取的主观性、高频词选择信息的损失以及词不同位置对词的重要性的影响的忽略等。

近几年来把文本和文献计量学的引用分析结合在一起的融合方法开始逐步受到学者的关注。这种融合的思想来自于信息检索领域。传统的搜索引擎主要是基于查询词与查询内容的匹配表明内容的相关性，然而实践证明单纯基

于查询词来进行结果的输出效率非常低。Page和 Kleinberg 分别提出的 PageRank和 HITS算法改变了传统的信息检索算法<sup>[6-7]</sup>,对检索的结果利用页面的链接信息赋予权重,从而把更加相关的信息放在检索结果的前面。由于引用和网页链接关系的相似性,信息检索把链接和词匹配相结合的研究也引发了文献计量学领域学者的研究兴趣。

2005年, Glenisson, Ganze和 Persson利用全文文本挖掘方法和文献计量学参考文献平均出版年结合在一起,对2003年ES的19篇会议论文的结构进行初步分析<sup>[8]</sup>,证明了这种方法的有效性。同年,他们又利用相同的方法扩大了数据集,对2003年《科学计量学》的论文进行了分析,作者提出利用全文比起利用标题和摘要结合的方法知识结构划分更加准确<sup>[9]</sup>。

2006年, Janssen等人利用文本挖掘方法得到的词文献矩阵和利用文献耦合得到的参考文献和文献的矩阵,分别利用相加求平均值和逆卡方法,把从两个不同角度得到的文献不相似矩阵利用统计方法结合起来,证明了这两种方法都很好地提高了领域主题分类或者知识结构的划分绩效,分类的结果更加准确。与 Janssen等人的方法不同, Glenisson的方法是从两个角度分别分析了同样的数据,利用引用数据来证明全文文本挖掘的有效性,而 Janssen的方法是把两个来源的数据在分析中同时进行考虑,把两个数据的特征进行了融合,证明了方法的有效性。2007年, Janssen的博士论文中对把文本挖掘和文献耦合方法的统计结合进行详细的阐述<sup>[10]</sup>,并在图书馆和信息科学领域以及生物信息学领域进行了实证的分析。这篇论文是目前利用文本挖掘和文献耦合两种数据源进行信息融合方法研究的代表。2009年,以其为代表的研究团队又利用这种方法根据期刊数据集对ESI的分类和ESI的心理学教育学领域的知识结构划分进行了研究<sup>[11-12]</sup>。

2.1.3 基于共词和共引结合的方法研究 共包括3种方法。

1) 基于共引基础的共词和共引结合方法。1988年, Mullins在《科学论文的结构分析》一书中就提到,要从科学论文的每一个方面来展开研究,从标题到参考文献,从图表到写作风格以及词的利用,分析论文的每个方面都能得到有价值的信息<sup>[13]</sup>。1991年, Braam等人利用共引分析得到的聚类作为基础<sup>[14-15]</sup>,利用引用这些聚类文献引用的标题词来进行共词分析,把共词分析和共引分析结合在一起,不但有利于清晰地标注类名,而且把二者结合起来有利于提高对学科结构研究的全面性和深入性。共引聚类可以反映专题发展的历史,而引文的内容词则能够更好地反映研究的现状。我国学者柴省三在1996年把这种方法作为一种新的理论和方法进行了介绍<sup>[16]</sup>,并在1997

年进行了应用研究<sup>[17]</sup>。侯跃芳等人利用这种方法对“妊娠糖尿病”医学领域的发展以及这种方法的可靠性进行了深入的研究<sup>[18-19]</sup>。

2) 基于词汇引用图。Jo等人基于在文献引用图中联系紧密的文献主题更相关这样的假定<sup>[20]</sup>,提出利用词汇引用图关系来抽取文献的主题,从而为主题探测打下很好的基础。对于给定的词汇A,假设H<sub>1</sub>是A和主题相关,假设H<sub>0</sub>是A和主题不相关。观察A的词汇引用图O(G<sub>A</sub>)在假设H<sub>1</sub>条件下的概率的对数值,和假设H<sub>0</sub>条件下O(G<sub>A</sub>)的概率的对数值,利用两个条件概率的差来表示和主题A是否相关,如公式(1)所示。

$$\begin{aligned} \text{TopicScore}(A) &= \lg P(O(G_A) | H_1) - \lg P(O(G_A) | H_0) \\ &= \lg \left[ \frac{P(O(G_A) | H_1)}{P(O(G_A) | H_0)} \right] \end{aligned} \quad (1)$$

作者利用这种方法在Citeseer和ArXiv中进行了验证,证明了这个方法的有效性。

吴清强也采用了这种词汇引用图方法来确定词汇的主题相关度<sup>[21]</sup>,从而确定知识结构分析中的重要词汇,作者也提出了检验词T的主题相关性方法:假设H<sub>1</sub>表示T与数据集研究主题相关,是主题词汇。假设H<sub>0</sub>表示词汇T与数据集研究主题不相关,不是主题词汇。在假定H<sub>1</sub>条件下的词汇的概率为P(T|H<sub>1</sub>),在假设H<sub>0</sub>下词汇的条件概率P(T|H<sub>0</sub>),比较两者的大小就可以判断词汇T偏向是主题词还是偏向非主题词,如公式(2)所示。

$$PT = P(T|H_1) - P(T|H_0) \quad (2)$$

由此可见,两个作者提出的方法其实是相同的,只不过用于分析的目的不同。从吴清强的论文来看,词汇引用图是在文献引用图的基础上构建起来的,而文献引用图实际上是文献之间的互引关系图。这两个研究表明,在利用引用对数据集收敛可起到很好的知识结构划分的效果。

3) 基于引用为背景的共词分析融合方法研究。文献[22]提出了一个词-参考文献共现来进行研究主题影射的方法。作者指出一个领域或者专业的研究者共享的知识基础,可以从参考文献的选择反映出来。一个研究领域或者专业可以用包含研究问题和方法的期刊网络和参考大量交叉的文献集合来定义。用于划分知识结构的方法无论是共词和共引都存在不足,共引方法由于受到时滞的影响,反映的是学科的历史结构,而共词由于词的模糊性及特征表示方法的不足,得到的知识结构可能根本没有实际的意义。基于此,提出了一个不同的方法,把词和参考文献结合在一起进行分析,把科学领域看作是一个交流网络,利用科学出版物来勾画这些交流系统。这个思想是:研究者在进行研究工作时,会同时选择描述研究主题的词

和参考具体的文献。参考文献是词具体含义的背景信息，有了背景信息的词分析就减少了词本身的模糊性，把与文献相关的两个属性结合在一起，来共同决定研究专业的细粒度结构。利用这种方法在信息科学领域进行实证分析，从勾画的每年的知识结构图来看，知识结构的划分非常清晰。这种融合方法与基于共引基础上再进行共词分析方法的融合不同，它是在分析词关系的时候同时考虑了引用关系，但是是把引用作为背景信息来进行处理的。

从几种信息融合的方法来看，有的方法是对原始数据集的扩大，以期能够从更多的数据当中挖掘更加全面的领域知识结构。有的是在一种分析的基础上，把另一种方法作为补充或者背景信息来进行处理，这种方法没有改变数据集的数目，但也不是把特征进行融合，只是从表层把特征结合在一起来进行分析。有的方法是在对数据集收敛的基础上进行分析，例如词汇引用图方法，利用数据集文献之间的互引关系缩小分析的数据，然后针对这些数据进行主题探测和主题词的识别。还有一种视角独特的融合方法，就是把从不同特征得到的数据利用统计方法进行融合，这种方法从更深层次把数据真正地整合为一体来进行整体分析。目前国内在这方面的研究特别是从计量学角度出发的研究还非常少，是未来值得关注和探讨的问题。

## 2.2 不同的知识结构划分方法研究

从领域的知识结构划分来看，主要包含数据的准备—数据的特征提取—数据的标准化处理—数据的分类或者聚类等几个阶段，最后根据聚类或者分类的结果进行知识结构划分的分析。对于知识结构的划分方法可以根据不同标准分为以下几类：根据是否加入了时间维度可以分为静态的知识结构和动态的知识结构；根据采用的知识结构划分方法分为监督学习方法、半监督学习方法和非监督学习方法，每种学习方法又可以根据参数和规则分为很多不同类型的方法；根据知识结构是非交叉分为软划分方法和硬划分方法。本文重点研究非监督学习方法的不同类型，特别是硬聚类方法。

2.2.1 常用的知识结构划分方法 在领域知识结构划分方法研究中，最常用的包括多维尺度分析方法、层次聚类方法和主成分分析方法。White和 Griffiths在1981年提出共被引分析方法的时候，就采用了多维尺度分析方法和主成分方法对信息科学的知识结构进行了研究。后来的研究者也都沿用了这些方法。层次聚类分析一般采用树形结构将知识结构展示出来，结构关系非常的直观清晰，这也是文献计量学研究比较青睐的一种知识结构划分和可视化表示方法，这种方法特别适合聚类比较小的数据集，而且具有较高的聚类精度。但是MDS和层次聚类方法都存在可处理的数据有限的缺点，不适合处理大的数据量。目前

由于MDS和层次聚类方法处理数据能力的限制，很多知识结构划分研究都取了很高的阈值，但是这种只取少数高频或者高被引对象的方法只能识别领域的主要结构，这些主要结构的识别可能没有现实指导意义，无法实现新兴主题的探测，发掘潜在研究领域等重要的科学发现功能。

2.2.2 基于切分的知识结构划分方法 K-means算法是切分算法中运算效率最高，并且可以处理大量数据的一种聚类算法。它首先根据用户选择的聚类数量 $k$ 随机把数据划分成 $k$ 类，再计算类内的点到聚类中心的距离，然后根据距离不断的进行调整，直到所有点的平均方差最小才停止划分。这种方法简单、易懂，而且计算的复杂性不大，可以在很多研究中都证明其有效性。但是K均值算法也有其不足，主要是因为初始聚类中心的不确定性，所以聚类的结果不稳定，可以采用限定初始聚类中心的方法来提高算法的稳定性。由于对于领域知识结构的不可知，所以在操作上聚类数量的选择带有较大的随意性，太多的聚类数量增加了分析的难度，而太少的聚类数量分析结果又过于概括，无法识别有意义的信息。基于切分的知识结构划分方法还包括k-medoid方法、DBScan等方法，但是比起K均值方法来说不及其应用的广泛。

2.2.3 基于图的知识结构划分方法 近几年来，随着复杂网络研究的逐步深入，人们发现现实世界中的很多事物及其事物的关系都可以采用网络的结构表示出来。例如文献可以作为图的节点，文献之间的词相似关系或者引用关系可以作为文献的边，这样就可以构建文献关系网络。传统的基于图的子网络切分方法有很多，这里主要介绍近几年研究比较热的网络的社团结构划分方法。社团，community也被翻译作社区、子图，类似于聚类当中的类或者簇，是对大的网络结构划分得到的子结构。对于社团结构划分的方法研究非常的多，其中最知名的研究是Newman在2004年提出利用 $Q$ （模块度）来快速地划分社团结构的方法<sup>[21]</sup>。这种方法不但成为很多后续算法的基础，而且成了很多网络社团结构划分好坏的一个评价标准。这种基于图的知识结构划分方法可以处理大量的节点，而且算法的效率也相对比较高，主要还是基于网络的连接度来进行结构的划分方法，这种方法目前在知识结构研究中应用的还相对比较少。

2.2.4 聚类融合的知识结构方法 对于知识结构的划分好坏有很多的影响因素，包括知识的表示、数据的标准化方法以及聚类的算法。每种方法都有其不足，单纯采用一种方法似乎让人很难信服，于是聚类融合方法就应运而生。这种方法的思想是把同一种算法的不同参数或者不同算法的结果合并在一起，是聚类分析领域最近几年才开始出现的研究方法。聚类融合算法比起单一的聚类算法能够

得到更好的结果,而且方法具有很好的稳定性、并行性和可扩展性<sup>[24]</sup>。

从目前的知识结构划分方法来说,方法非常繁多,不同的方法在不同的应用中有不同的效果,很难说哪一种方法是最好的方法。虽然目前聚类融合和软聚类算法是研究的一个重点和趋势,但并不一定这种方法在每种应用中都适用。从目前的知识结构划分方法来看,应该尽快接受更新的方法,不能限于传统分析方法的套路原地踏步。

### 3 结束语

本文对各种层次的文献内容和引用特征融合方法进行全面的剖析和解读,并对相关的知识结构方法也进行了大致的概括和总结。知识结构划分方法作为文献计量学主要研究内容之一,从更加全面和更加科学的角度来进行研究是推动该领域研究发展的必然趋势。从内容和引用融合的角度考察文献以及文献相关的领域,一方面符合科学交流的多途径和多表现形式;另一方面也符合事物之间以及事物的内部之间是互相联系的哲学观点,有利于更加准确地进行知识结构的划分研究。但是目前来看,相对大量的知识结构划分研究而言,这些研究还显得非常的薄弱,需要不断深入地去研究。□

#### 参考文献

[1] JANSSENS F. Clustering of scientific fields by integrating text mining and bibliometrics [D]. Faculty of Engineering KU Leuven (Leuven, Belgium) 2007.

[2] WHILEH D et al. Author cocitation—A literature measure of intellectual structure [J]. Journal of The American Society for Information Science 1981 (32).

[3] SYNNESTVEDT M B. Data Preparation for biomedical knowledge domains visualization: a probabilistic record linkage and information fusion approach to citation data [D]. [S. l.]: Drexel University 2007.

[4] EOM S. All author cocitation analysis and first author cocitation analysis: a comparative empirical investigation [J]. Journal of Informetrics 2008 (2): 53-64.

[5] ZITTM et al. Delineating complex scientific fields by an hybrid lexical-citation method: an application to nanosciences [J]. Information Processing and Management 2006 42: 1513-1531.

[6] History of PageRank [EB/OL]. <http://PageRank.ws/2008/11/23/history-of-Pagerank>

[7] KLEINBERG J. Authoritative sources in a hyperlinked environment [J]. Journal of the ACM 1999 46 (5): 604-632.

[8] GLENNISON P et al. Combining full text analysis and bibliometric indicators: a pilot study [J]. Scientometrics 2005 63 (1): 163-180.

[9] GLENNISON P et al. Combining full text and bibliometric information in mapping scientific disciplines [J]. Information Processing and Management 2005 41: 1548-1572.

[10] JANSSENS F et al. Towards mapping library and information science towards mapping library and information science [J]. Information Processing and Management 2006 42: 1614-1642.

[11] ZHANG L et al. Hybrid clustering analysis for the domain of psychology sociology & education [M]. [S. l.]: Proceeding of ISSI 2009.

[12] ZHANG L et al. Hybrid clustering analysis for mapping large scientific domains [M]. Proceeding of ISSI 2009.

[13] MULLINS N et al. The structural analysis of a scientific paper [R] // VAN RAAN A F J (Ed). Handbook of Quantitative Studies of Science and Technology. New York: Elsevier Science, 81-105.

[14] BRAAM R R et al. Mapping of science by combined cocitation and word analysis: structural aspects [J]. Journal of The American Society for Information Science 1991 42 (4): 233-251.

[15] BRAAM R R et al. Mapping of science by combined cocitation and word analysis: dynamic aspects [J]. Journal of The American Society for Information Science 1991 42 (4): 252-266.

[16] 柴省三. 引文——内容词分析研究科学结构的最新理论与方法 [J]. 国外情报科学, 1996 (3): 58-62.

[17] 柴省三. 内容词——共引聚类分析及其在科学结构研究中的应用 [J]. 情报学报, 1997 16 (1): 69-74.

[18] 侯跃芳, 等. 应用引文共引聚类——内容词分析法对学科发展的研究 [J]. 情报学报, 2007 26 (2): 309-314.

[19] 侯跃芳, 等. 引文——内容词分析法反映专题学科发展历史及现状的可靠性分析 [J]. 中华医学图书情报杂志, 2007 16 (3): 58-62.

[20] JO Y et al. Detecting research topics via the correlation between graphs and texts [R] // KDD-2007 Proceeding of The Thirteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2007: 370-379.

[21] 吴清强. 主题结构构件技术优化研究 [D]. 北京: 中国科学院文献情报中心, 2008.

[22] BESSELAAR P V D et al. Mapping research topics using word-reference co-occurrences: a method and an exploratory case study [J]. Scientometrics 2006 68 (3): 377-393.

[23] NEWMAN M E J et al. Finding and evaluating community structure in networks [J]. Physical Review E 2004 69.

[24] 阳琳, 等. 聚类融合方法综述 [J]. 计算机应用研究, 2005 (12): 8-10.

作者简介: 任红娟, 1979年生, 博士生。  
张志强, 1964年生, 教授, 博士生导师。  
收稿日期: 2009-12-16