情报学报 ISSN 1000-0135
第32卷第9期900-911，2013年9月
JOURNAL OF THE CHINA SOCIETY FOR SCIENTIFIC AND TECHNICAL INFORMATION ISSN 1000-0135
Vol. 32 No. 9, 900-911 September 2013

doi:10.3772/j.issn.1000-0135.2013.09.001

基于NEViewer的学科主题演化可视化分析[1)]

王晓光 程齐凯

（武汉大学信息管理学院，武汉 430072）

摘要 为了开发更加准确高效的学科新兴趋势探测方法，必须加强科研主题演化规律的研究。本文提出了一种新的基于共词网络社区演化分析的研究框架。我们基于社区主题表示算法和社区相似度匹配算法，构建了一个科研主题演化分析模型，并开发了一款新颖的网络社区演化分析软件NEViewer。与已有的科学图谱分析软件相比，NEViewer的创新在于：(a)设计一套时序网络社区演化分析框架；(b)实现了多个网络社区演化分析算法；(c)以冲积图和赋色网络图两种创新性的方式揭示了网络社区演化的宏观过程和微观细节。利用NEViewer对中文计算机学科进行的实验结果表明NEViewer在复杂网络社区演化可视化分析上是可靠的和有效的，借助共词网络进行学科主题演化研究的思路也是可行的。

关键词 新兴趋势探测 共词网络 网络演化 社区识别 可视化 NEViewer

Analysis on Evolution of Research Topics in a Discipline Based on NEViewer

Wang Xiaoguang and Cheng Qikai

(*School of Information Management, Wuhan University, Wuhan* 430072)

Abstract The evolution rules of research topics in a discipline is the key to develop new emerging trend detection methods. This paper proposes a new research frame based on co-word network evolution analysis. The knowledge structure of a discipline can be expressed by a special co-word network communities in that network mean topics. A topic evolution analysis model is created based on community match mechanism. NEViewer presents three key features that are remarkable compared to other science mapping software tools: (a) a powerful analyzing module within a longitudinal framework; (b) the use of several network community evolutions analyzing algorithms; (c) revealing the macroscopic shifts and microcosmic details of evolution based on alluvial diagram and colored network. The result from an experiment within Chinese computer science field showed that NEViewer is effective and liable. The research process, using co-word network analysis research topics evolution in disciplines, is also feasible.

Keywords emerging trend detection, co-word network, network evolution, communities detection, visualization, NEViewer

1 引 言

学科新兴趋势识别是近年来情报学的研究热点之一[1~7]。随着复杂网络理论和信息可视化技术在情报学领域的深入应用，知识图谱研究得到了快速发展，Leydesdorff、Boyack、Börner、Chen等在该领域取得了一系列令人注目的成果[8~11]。相关的成果

收稿日期：2012年10月18日

作者简介：**王晓光**，男，1978年生，副教授，主要研究方向：知识网络、语义出版。E-mail：whu_wxg@126.com。**程齐凯**，男，1989年生，博士研究生，主要研究方向：信息检索、文本挖掘。

1) 本文受国家自然科学基金项目“基于语义共词网络演化的学科新兴趋势浮现机理与探测研究”(71003078)和“知识网络的形成机制及演化规律研究”(71173249)资助。

表明知识网络是一种重要的情报学研究工具。知识网络的复杂结构与学科领域之间天然的对应关系使得知识网络挖掘在潜在知识发现、学科热点和学科新兴趋势探测上具有重要的方法论价值[12]。共词网络是知识网络中的一种形式，Börner、Chen 和 Boyack 等的研究都表明共词网络不仅可以作为绘制知识图谱的基础[10~12]，还与引文网络和共被引网络一样具有重要的方法论价值，而且相比较而言，将共词网络用于新兴趋势探测较之传统的探测方法还有着一定的优势[13,14]。

新兴趋势探测一般包括三个环节：主题表示（representation）、主题识别（identification）和主题判定（verification）。基于以往的科研经验可知，判定一个主题是新兴主题、热门主题、衰退主题还是死亡主题，都必须考虑时间维度，即在考虑主题自身科学价值的同时，也必须考虑该主题及其相关主题在已往的表现情况，即是否出现过、何时出现以及近几年的发展情况。任何一个学科领域的科研主题都不是凭空出现的，很多主题都是在已往的知识基础（Knowledge base）上孕育而生的，所以判定一个学科领域中的新兴主题，必须要了解该领域中所有主题的演化历史及当前的态势。

为了分析一个学科领域的研究主题的演化过程，我们开发了一款新颖的复杂网络演化可视化分析软件 NEViewer。该软件不仅可以实现主题表示、主题识别和主题判定，还能以可视化的方式展现科研主题演化的宏观过程和微观细节。接下来，文章将介绍 NEViewer 的基本框架和相关算法，此后展示一个利用 NEViewer 进行的实验及实验结果，最后文章对 NEViewer 的价值、意义和不足进行了总结。

2 方法框架

科学研究的成果常以学术文献的形式呈现，这些文献中的关键词可以被视为文献的“指纹”[15]，代表了学术文献的研究主题[13]。利用文献中的关键词共现关系可以构建共词网络，这样的共词网络揭示了特定学科领域的科研主题及其相互关系。不同时间段上的共词网络存在差异，这种变化形成了共词网络的演化过程，同时也揭示了研究主题的演化。

基于以上思想，我们把主题演化分析的流程设置如下：首先，对原始的科学文献进行处理，得到一系列时序共词网络。然后，通过社区发现算法找出每个时段上的网络社区，并为每个社区赋予主题标识。接着利用相关性算法，探测前后时段中网络社区间的相关性，以此确定社区演化关系，最后利用信息可视化方法展示这一过程。整个流程如图 1 所示。

2.1 共词网络的构造

为了构建共词网络，我们首先做如下定义：时序文档集 $D=\{D_1,D_2,\cdots,D_n\}$，D_t 表示时间段 t 内刊载的文档构成的文档集合，$D_t=\{d_{t1},d_{t2},\cdots,d_{tm}\}$，$d_{ti}$ 表示 t 时间段内编号为 i 的文档，$d=\{w_1,w_2,\cdots,w_z\}$，w_k 为文档的第 k 个关键词；对应于 D，$G=\{G_1,G_2,\cdots,G_n\}$，t 时间段的共词网络 $G_t=\{V,E\}$，其中，V 为网络的节点集合，E 为连接集合。

本文使用学术文献中的关键词作为学科主题的表征，以此构建共词网络。词汇网络关系的定义规则如下：

（1）词汇 w_a 和 w_b，如果 $w_a\in d$ 且 $w_b\in d$，则认定 w_a 和 w_b 存在着一次共现关系，此处不对关系进行加权，因为某些文档的词汇的强共现关系会对最后的分析结果造成干扰。

（2）如果 w_a 和 w_b 在 n 篇文档中共现，则 w_a 和 w_b 间存在着权值为 n 的联系。

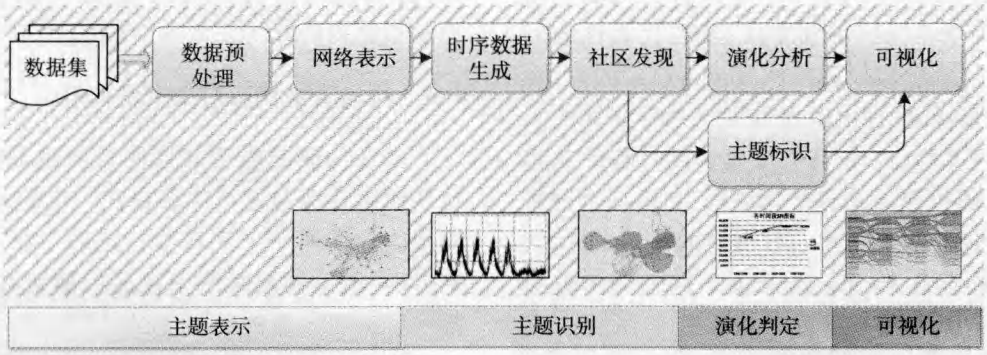

图 1 基于共词网络的科研主题演化分析框架

根据这一规则，给定 D_t，网络 Gt 的构造方法如下：

(1)构造一个空的共词网络 G；

(2)遍历文档集 D 中的文档，对于每一个文档 d，其概念描述性词汇 $W=\{w_1,w_2,\cdots,w_n\}$，对于任意 w，如果 w 没有在 G 中出现，将 w 作为一个节点加入 G；对于 W 中的任意词汇组合 w_aw_b，如果 G 中 w_a 和 w_b 之间没有建立联系，构建两者之间的联系，设置关系权值 $l_{a,b}$ 为1，否则将令。

为了得到时序共词网络，需要根据时间片对文档集进行划分。TimeLine 方法和固定时间窗口是两种常见的划分方法[16~18]。TimeLine 方法复杂度较高，划分效果也难以得到保证。因此，本文使用固定时间窗口的方法：设定长度为 tl 的时间段为一个时间窗口，将文档集 D 划分为 n 份，对每份文档数据分别构建共词网络，得到共词网络序列 G。

2.2 共词网络中的社区发现

在信息科学领域，社区发现存在两个方向：一种是基于拓扑关系的社区发现，另一种是基于主题的社区发现[19]。基于拓扑关系的社区识别主要依赖图论方法，基本的假设是社区内部的关系应该比社区之间的关系更加紧密。这种识别方法主要依赖于网络拓扑关系，而不考虑网络节点和网络关系的性质，因而适合于任何复杂网络，如共被引网络、共词网络。基于主题的社区识别主要针对那些网络节点是一个或多个文本集合的网络，如博客网络和大学网络。这种识别主要依赖两个节点间拥有的主题相似性，在层次聚类基础上形成一个树状结构，以此显示哪些节点属于一个社区。相对而言，基于拓扑关系的社区识别方法比层次聚类法更有优势，因为它不需要预先设定聚类数目和确定树状图的层次切割点。

目前，理论物理学界和计算机学界已经基于图论思想提出了众多社区识别算法，最有代表性的一类方法是基于优化网络模块度(Modularity)的方法。模块度是由 Newman 提出的衡量网络划分好坏的一种指标。模块度值，也叫 Q 值。模块度计算的基本思想是：完全随机网络没有社区结构，如果一个网络有良好的社区结构，就存在一种对该网络的划分，使得这种划分对应一个较高的 Q 值。对于真实世界的网络而言，Q 的取值一般介于0.3~0.7之间。基于模块度的方法大多旨在优化 Q 值，希望找到一种网络的划分，使得这种划分对应的 Q 值最大。

从本质上来说，基于模块度的算法是根据边的中介性和模块度的变化来进行社区识别的。提出模块度方法之初，该方法只能适用于无权网络。2004年，Newman 又提出了一个新算法，将模块度算法扩展到了加权网络上[20]。新算法与老算法的思想并无本质不同，只是在进行边切割的时候，新算法不仅考虑边的中介性还考虑它的权重。新算法的模块度计算公式为：

$$Q=\frac{1}{2m}\sum_{ij}\left(A_{ij}-\frac{k_ik_j}{2m}\right)\delta(c_i,c_j) \qquad (2)$$

其中，A_{ij} 是节点 i 和节点 j 之间的边权重，代表网络中的所有边的权重之和，K_i 代表节点 i 的度值，C_i 代表节点 i 所在的社区。如果 $C_i=C_j$ 则，否则。经过改造，新的模块度计算方法适合于所有的网络。

从很多现实的网络来看，社区重叠现象十分明显，即部分网络节点同时隶属于多个社区。在知识网络中，这种情况也非常常见，如一篇文献可能被多个学科领域引用，因而属于多个学科社区。一个概念在多个学科存在，分别代表不同的内涵，也会形成多重社区身份现象。如果一个算法能够允许社区重叠的话，无疑该方法的灵活性和精确性更加理想。Palla 等在2005年提出了一个基于 K-cliques 思想的允许社区重叠的社区识别算法[21]，并开发了一个社区识别与可视化软件 CFinder。Brian Ball 等基于生成式模型提出了一种类似于主题模型的社区识别算法，能够较好地处理社区重叠问题[22]。Lancichinetti 与 Fortunato 比较了多个方法之后，认为 Rosvall 和 Bergstrom 的基于信息论的算法性能最佳[23]，但是 Blondel 等[24]的算法也相当优秀。

2008年，McCain 利用引文网络和社区发现算法进行了科研主题的识别研究[25]。随后，Wallace、Gingras 和 Duhon 又利用两个案例研究证明了将社区发现方法用于研究方向的识别不仅是可行的，更是一种非常理想的思路，它能比传统的共被引分析揭示出更多的知识领域的结构细节[26]。

2.3 共词网络社区对应主题的表示

已往的研究表明共词网络内也存在社区现象，这些网络社区与学科体系存在一定的对应关系。不同层次的词汇社区代表了特定的学科、专业以及研究方向，所以共词网络中社区的演化在一定程度上揭示了科研主题的发展过程[13,14]。

在确认了共词网络内的社区具有特定的指示性意义后，下一步就必须确定这些社区代表的主题。

由于共词网络中的节点就是文章关键词，所以确定社区代表主题的过程也就转化为寻找社区核心节点的过程，少数核心节点代表了社区对应的科研主题。

在复杂网络中，节点的重要性指标有很多，除了传统的中心度、声望等指标外，还有 Pagerank 值。这些指标都从网络全局层面进行考虑，计算每一个节点在整个网络中的边数、中介性以及与其他节点的连接情况，进而判断出全局层面的核心节点。这些指标虽然能揭示出每个节点在全局范围的地位，但无法揭示一个节点在一个特定社区内的重要性。

为了寻找社区内的代表性节点，我们使用了 Z-Value 值。该指标由 Guimerà 等提出[27]，它可以衡量网络节点与其他节点联系的紧密性，是一个在地区层面而非全局层面揭示节点重要性的指标。

Z-Value 定义如下：

$$Z_i = \frac{k^i_{s_i} - < k^j_{s_i} >_{jes_i}}{\sqrt{< (k^j_{s_i})^2 >_{jes_i} - < k^j_{s_i} >^2_{jes_i}}}$$

其中，k^i_s 表示节点 i 到社区 s 中其他节点的连接数，s_i 表示 i 所在的社区，而 $< \cdots >_{j\in s}$ 表示社区 s 中所有节点的平均数。Z-Value 值越高表明节点与其所在社区内其他节点联系越紧密。

在使用 Z 值后，根据经验，共词网络中每个社区的代表节点，即对应的主题，就可以由一个或多个 Z-Value≥2.5 的节点表示[27]。

2.4 共词网络中的社区演化分析

在共词网络中，网络社区不是一成不变的。在不同的时间段内，网络社区的数量、大小、密度、结构等属性并不一致，所以网络社区的演化既包括社区自身内部节点、关系和结构的变化，也包括社区间关系和位置的变化。由于我们把每个网络社区视为一个研究主题，并重点关注研究主题的演化过程，所以我们参考 Palla、Barabási 和 Vicsek 的做法[28]，将网络社区的演化过程分为六种形式，分别是产生、消亡、分裂、合并、扩张和收缩。

定义 1：产生：指 t 时间段不存在的社区，在 $t+1$ 时间段产生；

定义 2：消亡：前 t 时间段存在的社区，在 $t+1$ 时间段没有存在；

定义 3：分裂：前 t 时间段的社区，$t+1$ 时间段分化成为两个或多个新的社区；

定义 4：合并：前 t 时间段的两个或者多个社区，在 $t+1$ 时间段合成一个新的社区；

定义 5：扩张：前 t 时间段存在的社区，在 $t+1$ 时间段继续存在，但规模扩大；

定义 6：收缩：前 t 时间段存在的社区，在 $t+1$ 时间段继续存在，但规模缩小。

这六种演化过程，均涉及 t 和 $t+1$ 两个连续时间段的社区关系，所以分析科研主题的演化过程就简化为为 t 时段的所有网络社区寻找前驱和后继。在学科主题演化中，存在着主题消亡的情况，并不是每个主题都有合适的后继。而大多数科研主题会存在前驱，所以在探测社区前驱后继关系时，我们采用自后向前为社区寻找前驱的方法。

寻找社区前驱和后继本质上是一个度量社区相似度的问题。本文假定：如果前后两个连续时间段中的社区相似度超过一定阈值，则两个社区存在演化关系。定义社区的前驱为：

$$Pre(M_{(t+1)j}) = (M_{ti} \mid M_{ti} \in G_t, d(M_{ti}, M_{(t+1)j}) < \delta) \cup \operatorname{argma}_{x_{M_{ti}\in G_t}} (d(LM_{ti}, LM_{(t+1)j})) \quad (1)$$

其中，δ 是可调节的阈值，d 是相似度计算公式。

相似度计算公式 d 的定义非常关键。在网络分析中，度量社区相似性有三种基本方法：节点重合度[28]，关系相似性[29]或者前两个指标的结合。本文主要使用节点重合度指标，同时也提出了一种新的结合节点重合度和关系相似性的度量指标 FS。

(1) 节点重合度的度量指标

节点重合度的基本度量指标有点积、余弦相似度、Jaccard 系数、广义 Jaccard 系数等。本文定义了一种加权匹配度指标用于度量社区的相似性。给定社区 M_x 和社区 M_y，各自对应的词汇集合为 C_x、C_y，加权匹配度定义为：

$$Sim(M_x, M_y) = \frac{\sum_{v\in C_x \cap C_y} W(v)}{\min(\sum_{v\in C_x} W(v), \sum_{v\in C_y} W(v))} \quad (2)$$

其中，$W(v)$ 表示节点 v 的频度，$\min(x, y)$ 为 x 和 y 中较小的值。

如果要求两个社区不但节点相似，且节点规模也相似，使用点积是一个很好的选择。但是，在主题演化分析中，社区与其前驱社区的节点规模可能相差较大，为了允许这种情况存在，可以使用余弦相似度或者广义 Jaccard 系数。二元属性数据的向量使用 Jaccard 系数更为简便。加权匹配度是最为稳定的指标，在大多数情况下都能较好地度量社区的相似性。

(2) 基于核心节点的重合度匹配

核心节点是那些在社区中具有较高重要性的节

点，中心性、声望、Z-Value 等都可以用于度量节点的重要性[30]。本文使用 Z-Value 度量节点在社区中的重要性，Z-Value 的定义见文献[27]。定义社区 M 的核心节点集合为：

$$H(M) = \{v | Z(v) > \delta, v \in M\} \quad (3)$$

$Z(v)$ 表示节点 v 的 Z-Value，δ 是一个人工设定的阈值，一般地，$\delta = 2.5$。

文献[17]认为社区的发展状态由核心节点决定，本文接受这一看法，定义社区 M_x 和社区 M_y 的相似度为：

$$HS(M_x, M_y) = \text{sim}(H(M_x), H(m_y)) \quad (4)$$

其中，sim 为两个节点集合的相似度计算方法，如广义 Jaccard 系数、余弦相似度等。

(3)结合节点重合度和关系相似性的度量指标

公式(2)～公式(7)都是节点重合度的度量指标，没有涉及对关系的度量。Berger[29]使用边关系度量社区的相似度，社区关系相似度通过公式(8)计算得到。

$$ES(M_x, M_y) = \frac{E(x) \cap E(y)}{E(x) \cup E(y)} \quad (5)$$

受 Berger[29]、吴斌[18]等的启发，本文提出了一个新的相似度计算指标，定义社区 M_x 和社区 M_y 的相似度指标 FS 为：

$$FS(M_x, M_y) = EJ(N_x, N_y) * HS(M_x, M_y) * ES(M_x, M_y) \quad (6)$$

其中，$EJ(N_x, N_y)$ 为节点重合度的度量公式，$HS(M_x, M_y)$ 为社区的核心节点重合度，$ES(M_x, M_y)$ 为社区的关系相似性度量。FS 指标兼顾了对节点重合度、关系重合度、核心节点重合度三个方面的度量，可以更准确的地反映社区的相似性。

2.5 共词网络社区演化过程可视化

在研究复杂网络演化时，Rosvall 曾借鉴地理学领域的冲积图(alluvial diagram)提了一种社区演化关系分析方法[31]，如图 2 所示。在冲积图中，矩形颜色块表示社区，两个时间段的矩形之间的曲线形色块表示演化的过程，颜色块的高度表示社区的节点规模。前一个时间段的主题同后一时间段的社区之间存在的演化关系通过色块的融合、分化加以体现：时段 t 的颜色块在时间 $t+1$ 分裂为两个或多个颜色块，表示社区在下一个时间段分裂为两个或者多个社区；时间段 t 的两个或以上的颜色块在下一个时间段发生融合，表示社区融合以及新社区的产生。

我们借鉴了冲积图的方法来表示主题的演化过程，但是该方法也存在一些不足：首先，原始的冲积图呈现方式并不能反映学科主题在当前时段的活跃度；其次，如果一个主题社区存在两个及以上的前驱，各个前驱在融合过程中的重要性并不能得到体现。因此，本文对 Rosvall 的方法做了一定的修改。

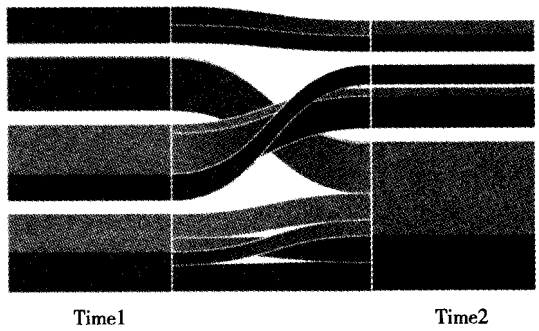

图 2 复杂网络中社区演化的冲积图表示

首先，为了体现不同学科主题在当前时段的活跃度，本文在构造冲积图时对学科主题进行排序，排序越靠前的社区越靠近图形的顶端。有两种排序方法可以使用：一是根据学科主题的词汇规模或者文档规模；二是将社区看做节点，计算社区在网络中的度值，根据社区的度值进行排序。

为了表现多个前驱在融合中重要性的不同，在绘制冲积图时需要将社区色块的颜色根据前驱的颜色设定。如果一个社区仅有一个前驱，则该社区的色块同其前驱的色块有着同样的颜色；如果社区有着两个及以上的前驱，则根据前驱各自的颜色为融合后的社区色块赋色：社区同前驱越相似，则社区的颜色越贴近该前驱的颜色。通过冲积图，我们可以非常直观地看到社区的演化情况，可如果要进一步地查看各个概念词汇的来源和去向，冲积图就无能为力了。

为了表现社区演化的细节，我们还设计了一种赋色网络图(图 3)，用于展示社区演化过程中的单个网络节点的来源和去向。

赋色网络图有后向赋色和前向赋色两种形式。后向赋色用于反映当前社区中节点在下一个时段的走向，前向赋色则用于表示当前社区节点的来源。如图 3 所示，Time1 时段的社区 A 在 Time2 时段有着两个后继 B 和 C，后向赋色根据节点分流情况的不同为 A 中的节点分别赋色。Time2 的社区 B 和 C，在 Time3 出现了融合现象，形成了社区 D，前向赋色根据 D 中节点的不同来源为节点赋不同的颜色。后向赋色和前向赋色的规则如下：

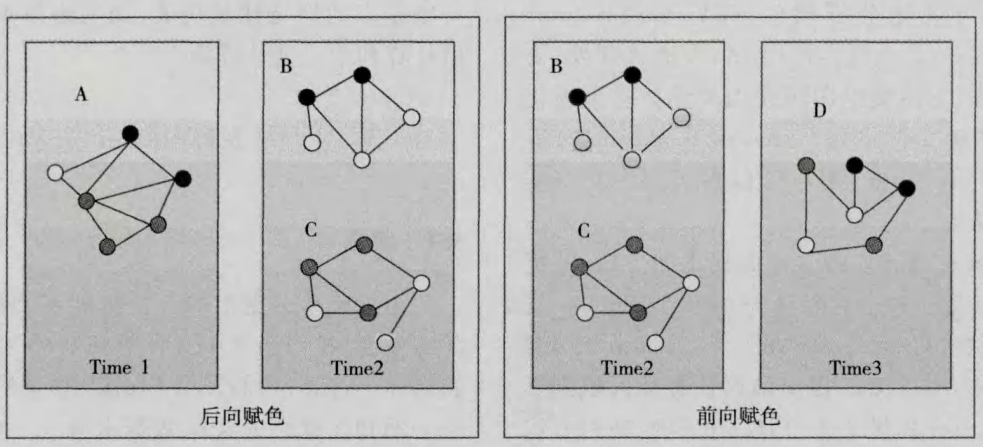

图 3　赋色网络图样例

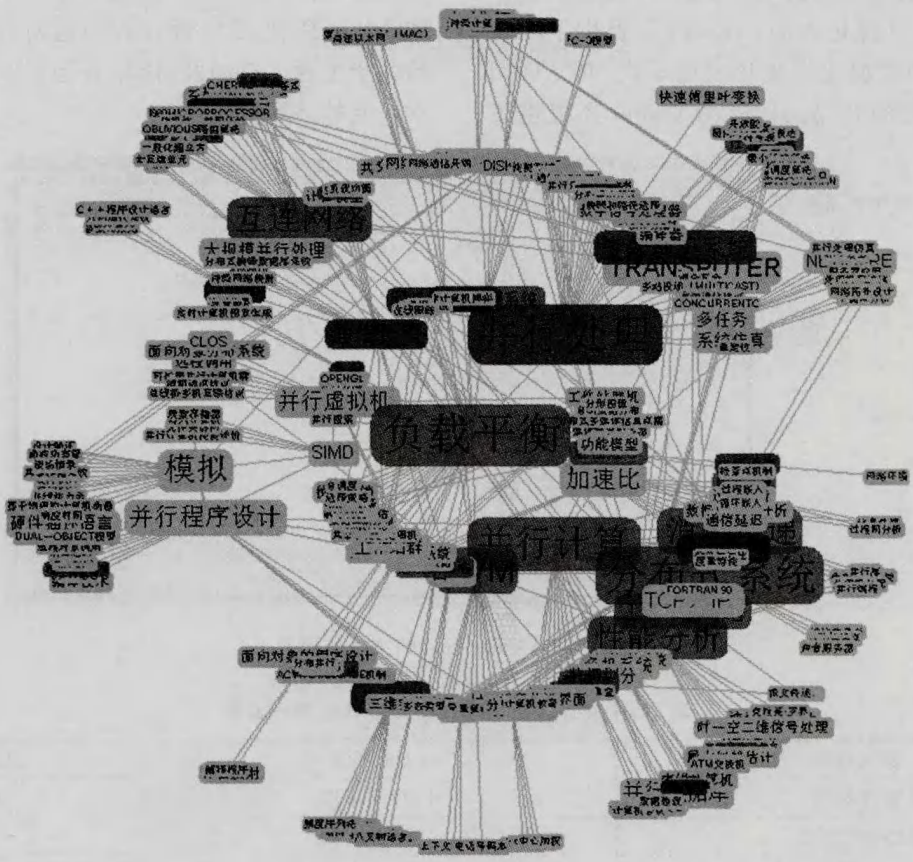

图 4　“分布式计算”研究主题的赋色网络层级布局样例

后向赋色：给定 t 时间段的社区 M_t，对社区 M_t 的任一词汇节点 v，如果社区 $cM_{t+1,i}$ 包含同样的词汇节点，令 VColor(v) = AColor($cM_{t+1,i}$)。其中，$cM_{t+1,i}$ 为社区在 $t+1$ 时段的第 i 个后继社区，AColor(M) 表示社区 M 在冲积图中的显示颜色，VColor(v) 表示 M_t 中节点 v 的颜色。

前向赋色：给定 $t+1$ 时段的社区 M_{t+1}，对 m_{t+1} 的任一节点，如果社区 $pM_{t,i}$ 包含同样的词汇节点，令 $VColor(v) = AColor(pm_{t,i})$。其中，$pm_{t,i}$ 为社区 m_{t+1} 在 t 时段的第 i 个前驱，AColor(m)、VColor(v) 定义同上。

为了表示不同节点在演化中的作用，赋色网络对节点分布使用了层级布局(hierarchical layout)，将核心的节点置于社区的中心位置，如图 4 示例。

2.6　NEViewer

基于上述分析框架和算法，我们开发了一款复

杂网络演化可视化分析软件 NEViewer(Network Evolution Viewer)。该软件使用 JAVA 语言开发,支持跨平台工作;支持美国印第安纳大学信息可视化实验室的 NWB 文件格式(Network Workbench File format);支持插件功能,用户可以参照程序文档实现自己的算法。

NEViewer 完整地实现了本文提出的方法,并提供了更多的功能选项。在社区识别上,NEViewer 支持 Blondel 等提出的 Louvian 方法[24]、Neman 的 MM 社区识别算法[32,33]、Ball 提出的可重叠社区识别算法[22]。NEViewer 提供了多个社区相似性判别方法的实现,包括 Jaccard 距离、广义 Jaccard 系数、余弦相似度、核心节点重合度、本文公式(6)给出的 FS 相似度指标;在可视化方面,NEViewer 提供了冲积图和赋色网络两种演化可视化表现形式,并允许用户定制赋色网络的布局;另外,NEViewer 还提供了一些基本的网络计量功能,如网络节点的 Pagerank 值计算和中心度计算等。

3 使用 NEViewer 进行实验

3.1 数据

为检验本文提出的分析框架和 NEViewer 软件的有效性,我们选择中文计算机科学领域进行了实验研究。首先,我们采用 CDBLP 收录的中文计算机核心期刊文献作为文献数据来源[34]。数据集包括 12 本期刊,37 847 篇文献,时间跨度为 1995 ~ 2010 年。分析使用的 12 本期刊都是中文计算机科学领域的核心期刊,文献质量较高且对计算机科学研究的各个主题都有涉及,能较好地反映中文计算机科学研究的进展情况。

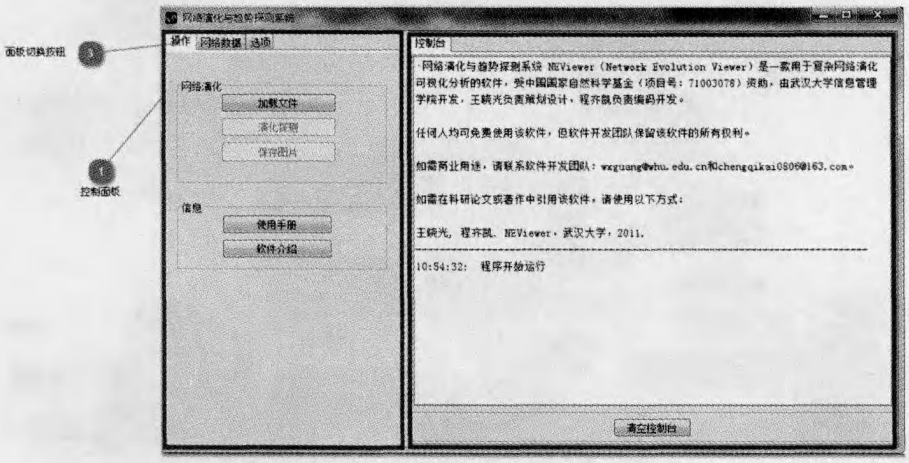

图 5 NEViewer 软件的主操作界面

表 1 中文计算机科学数据集期刊信息一览表

期刊名称	时间跨度	文献数量
软件学报	1995 ~ 2010	1 965
计算机学报	1995 ~ 2010	1 802
计算机研究与发展	1995 ~ 2010	2 293
计算机工程	1995 ~ 2010	11 388
中国图形图象学报	1996 ~ 2010	4 301
中文信息学报	1995 ~ 2010	1 209
计算机科学	1995 ~ 2010	4 664
小型微型计算机系统	1995 ~ 2010	3 183
计算机科学与探索	2007 ~ 2010	317
计算机辅助设计与图形学学报	1995 ~ 2010	1 867
电子学报	1995 ~ 2010	3 994
计算机科学技术学报	1995 ~ 2010	864
合计	1995 ~ 2010	37 847

表 2　共词网络的部分统计性质

网络统计指标	时间段			
	1995 ~ 1998	1999 ~ 2002	2003 ~ 2006	2007 ~ 2010
Nodes	7 524	14 360	23 714	27 891
Isolated Nodes	4	4	2	3
Edges	16 566	36 501	67 312	77 838
Mean degree	4.403 5	5.083 7	5.677	5.581 6
weakly connected components	635	742	837	955
largest connected component	701	2 227	4 551	5 341
Density	0.000 59	0.000 35	0.000 24	0.000 2

（Nodes 指节点数量，Isolated Nodes 为孤立点数量，Edges 为边的数量，Mean degree 为平均度，weakly connected components 为弱联通分量的个数，largest connected component 为最大联通子图的节点大小，Density 为密度）

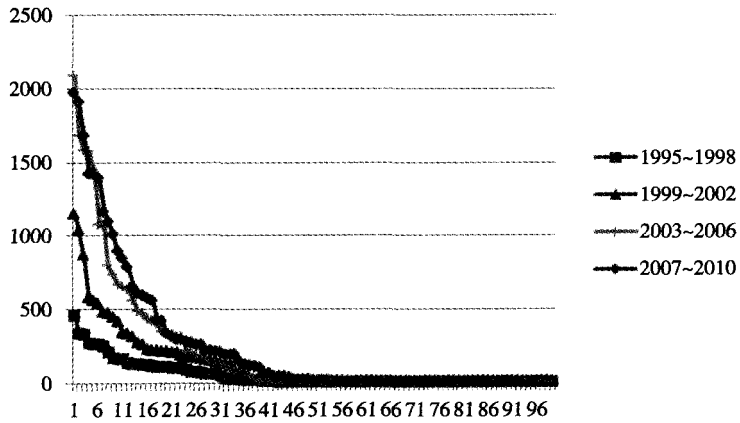

图 6　中文计算机科学各时段社区节点数分布图

为了构建时序共词网络，我们以四年为一个时间窗口，构建了四个共词网络，各个时间段共词网络的一些统计指标如表 2 所示。从基本的统计数据可以看出，1995 ~ 2010 年，中文计算机科学研究共词网络的节点规模越来越大，节点间的联系也越发密切。节点的平均度值有所增加，但网络的密度却逐渐变小。

3.2　研究主题识别

使用 Blondel 的社区划分算法对各时段的共词网络进行社区发现，得到共词网络的社区结构。图 6 给出了各时段共词网络的节点数量排名前 100 的社区节点数量分布图，其中，横轴表示社区在所在时间段的节点数量排名，纵轴表示社区节点数量。

共词网络中存在一些节点规模过小的社区，这些社区的产生往往是因为文献作者标注关键词不准确导致，构成了分析过程中的噪声信息。因此，在演化分析中去除了节点数在 10 以下的社区。节点数在 10 以上的社区的统计信息见表 3。

表 3　中文计算机科学领域各阶段共词网络社区数量统计

时间段	社区数量	词汇量	平均词汇量
1995 ~ 1998	49	5 144	104.98
1999 ~ 2002	60	11 601	193.35
2003 ~ 2006	61	20 451	335.26
2007 ~ 2010	58	24 081	415.19

3.3　主题演化可视化

不同的社区相似度测量指标有着不同的侧重点，本文使用公式(2)给出的加权节点匹配度，相似度阈值设定为 0.3，得到演化分析结果，如图 7 所示。

图 7 显示了四个时段的主题演化过程，每个社区对应的主题利用社区中 Z-Value 最大的节点（关键词）表示和标记。从图 1 可以看出以下现象：

(a)随着时间推移,学科主题越来越多;(b)绝大多数的研究主题都存在分裂现象,极少有合并现象,这反映了中文计算机科学领域研究主题的分化;(c)研究主题的收缩和扩张现象都存在;(d)部分社区没有后继,这意味着存在研究主题消亡现象。

3.4 NEViewer有效性分析

在图7中,我们根据网络社区的中心度进行了排序,越靠近顶端的社区中心度越高,由此可以看到"神经网络"、"数据库"等主题一直保持在靠近顶端的位置,这些研究课题也是计算机科学研究的基础性课题。分布式计算在1995~1998时段排序较高,但在随后的时段里排序有着较大幅度的下降,相反,标记为"小波变换"的图像处理学科主题排序上升较快,这同近十年来图像处理研究的快速发展是吻合的。

图8给出了"机器学习"单个研究主题的演化冲积图,其中只列出了1995~1998年"机器学习"主题在后续时段的演化情况。可以看到,社区在1999~2002年时间段有着三个后继,分别是标记为"神经网络"、"算法"、"VLSI"的三个社区,第三个社区在社区标记上不是很恰当,这个社区主要的内容是形式逻辑研究。三个后继反映了机器学习研究的三个发展方向,统计方法、算法改进和基于规则的机器学习方法。统计方法是近年来机器学习的主流,在后期有着较多后继主题,基于规则的机器学习研究近年来受到的重视程度逐渐降低,这一现实同冲积图反映的情形是一致的。机器学习方法在2002年以前主要以"神经网络"、"遗传算法"为主,而在第三阶段,支持向量机方法得到了广泛的应用,这一实际情况在网络演化图中体现为"支持向量机"构成了2003~2006年时段标记为"遗传算法"的社区的核心节点。

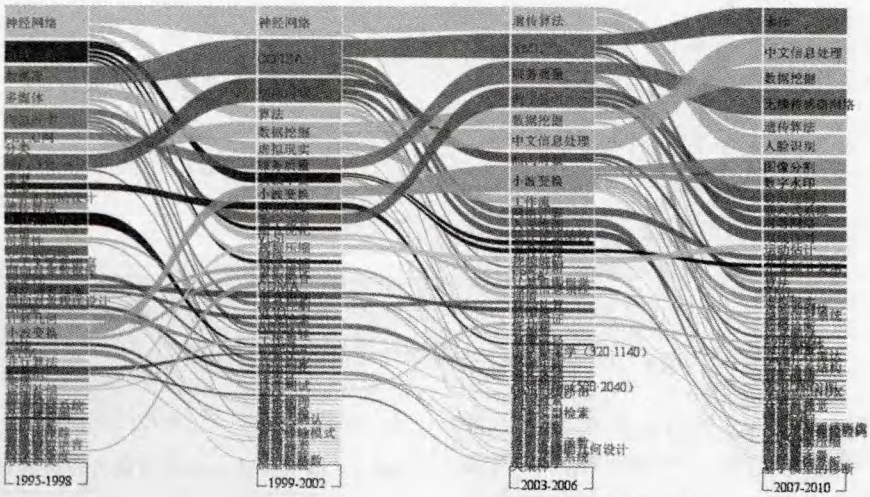

图7 中文计算机学科主题演化冲积图

图8 机器学习课题的演化路径

在图 8 中，2007 ~ 2010 年时段标记为“人脸识别”的研究主题构成了前一时段“机器学习”主题的后继。“人脸识别”研究属于图像处理范畴，那么，“人脸识别”主题是如何成为“遗传算法”主题的后继呢？图 9 给出了人脸识别课题的前向赋色网络，其中，“小波变换”是图像处理的传统方法，而支持向量机、特征提取等则是机器学习领域的概念，人脸识别综合应用了机器学习方法和图像处理技术。实际上，在绘制图 7 时本文设定的阈值是 0.3，如果将阈值下调为 0.2，就可以发现 2007 ~ 2010 年时段的“人脸识别”主题正好有着两个前驱，分别对应着“机器学习”和“图像处理”。

4 讨 论

新兴趋势探测研究是当前情报学研究的热点，为了发现学科新兴趋势，必须首先了解学科领域中的科研主题的演化过程、规律及态势。无论从情报学角度还是从复杂网络学角度来看，判断两个科研主题是否存在演化关系以及存在何种演化关系都是一个难题。从情报学角度来看，对于科研主题之间关系的判断需要结合特定的学科背景，深入分析这两个主题的学科定位、学术目标、发展历史、研究方法、研究人员等信息；从复杂网络学角度来看，对网络社区演化的判断需要详细分析两个社区的节点、边、结构等信息。即使两个社区拥有相似的节点、边和结构，判定两个社区之间的关系类型也不容易。在没有公认的判定标准的条件下，以相似度作为社区演化关系的判定标准是最为可行的方法。但在此过程中，如何设定阈值又是一个难题。阈值的大小将直接决定社区演化关系的判定，对于不同的网络是否设定相似的阈值都必须在实践过程中摸索确定。

目前，已经有一些用于知识图谱绘制和科研主题演化分析的软件，如 CiteSpace、VosViewer、Network Workbench、SciMat 等。与这些工具相比，NEViewer 更注重演化分析，它不仅实现了我们提出的多种网络社区演化分析指标，还以新颖的冲积图和赋色网络图的形式可视化地展示了网络社区的宏观演化过程和微观演化细节。

我们的实验表明网络社区的六种演化状态——产生、消亡、分裂、合并、扩张与收缩都是存在的。在

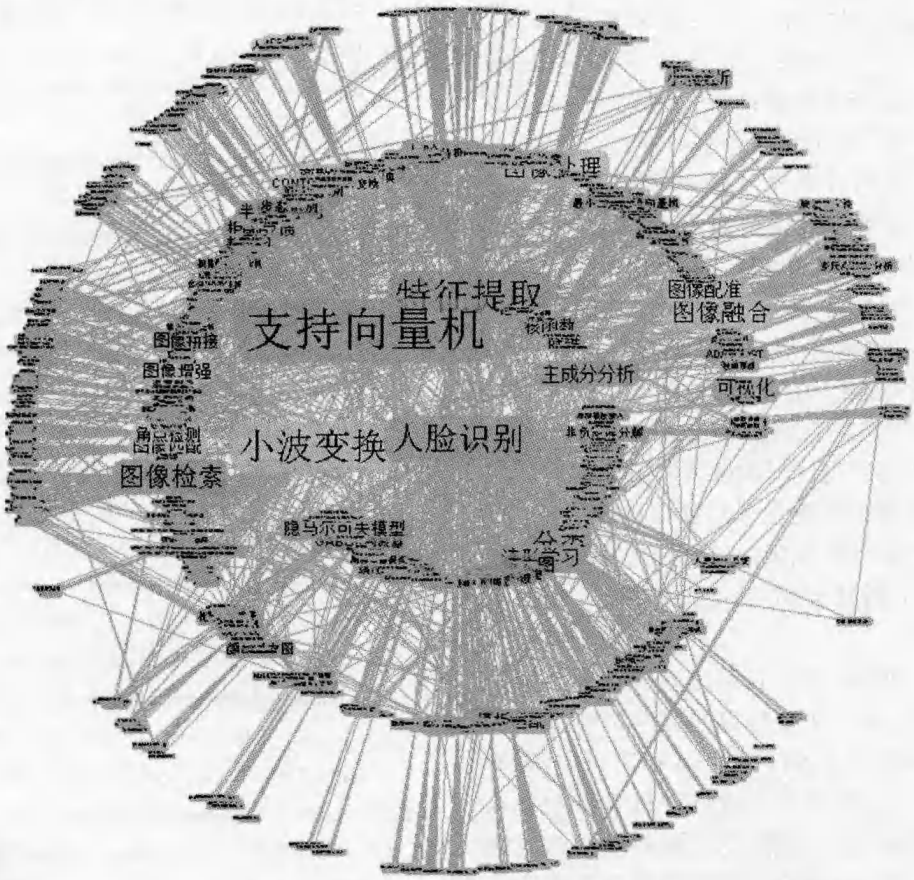

图 9 “人脸识别”主题的前向赋色网络

六种形式中,分裂和合并都导致了新主题的出现,但在这两种形式之外,某些时候,主题会直接得以产生。分裂和合并较主题直接产生更容易出现,直接产生的主题往往是突发点所在。分裂和合并都是有一定前兆的,探索这种前兆及其背后的诱因,可以对新主题的出现起到预测作用。扩张意味着主题的研究内容外延的扩大,收缩则反映了学科主题研究的衰落。消亡是主题持续收缩的结果。这些演化状态在新兴趋势探测中有着重要的意义。

5 总 结

为了分析科研主题的演化规律,本文提出了一种新的基于共词网络的科研主题演化分析框架,给出了共词网络中社区对应主题的表示算法以及社区演化关系的判断算法,并以此为指导,开发了NEViewer软件。与以往的科研主题演化分析思路不同,本文提出的分析框架不强调词频的变化,而是将复杂网络和网络演化思想引入情报分析过程,强调词间关系的变化,试图在网络视角下发现科研主题演化特征。

NEViewer在框架设计上具有很好的扩展性。NEViewer工作的四个流程:主题表示、主题识别、演化判定和可视化是相对解耦的,每一个步骤都可以在算法上进行扩展,如在主题表示步骤上,可以使用主题模型对词汇进行预处理,归并相似词汇,排除无意义词汇,这样做并不影响后面步骤的正常进行。NEViewer已经为算法扩展预留了接口。

与已有的类似软件相比,NEViewer的创新之处在于:(a)设计一套时序网络社区演化分析框架;(b)实现了多个网络社区标识和演化判别算法;(c)以冲积图和赋色网络图两种创新性的方式揭示了网络社区演化的宏观过程和微观细节。作为一款新颖的复杂网络演化可视化分析软件,NEViewer除了可以应用在信息计量学研究和科研决策管理上之外,也可以用于分析社会网络、企业网络、人际网络等多种复杂网络。

最后,本文提出的新型科研主题演化分析框架还存在一些问题,如本文在构建共词网络时没有对关键词的角色进行区分,有的代表了研究对象,有的则代表了研究方法,将所有关键词同样看待会导致主题演化分析准确度的下降;还有我们使用的网络社区演化关系判断方法较为粗略,没有考虑网络结构相似性。但不管怎样,与传统的基于词频的分析思路相比,本文提出的基于共词网络社区演化分析的框架还是开辟了一条新的研究路线。在未来工作中,我们将重点关注文章关键词的归一化问题,以此弥补该研究框架在主题判断精度上的不足。此外,我们也将进一步改进NEViewer的算法效率,并提高该软件易用性。

参 考 文 献

[1] Tan A H. Text mining: The state of the art and the challenges. Proceedings of the PAKDD 1999 Workshop on Knowledge Discovery from Advanced Databases[C]. 1999: 65-70.

[2] Wang J, Xu C, Li G, et al. Understanding research field evolving and trend with dynamic Bayesian networks[J]. Advances in Knowledge Discovery and Data Mining, 2007(4426): 320-331.

[3] Moerchen F, Fradkin D, Dejori M, et al. Emerging trend prediction in biomedical literature: AMIA Annual Symposium Proceedings[C]. 2008: 485-489.

[4] Schiebel E, Hörlesberger M, Roche I, et al. An advanced diffusion model to identify emergent research issues: the case of optoelectronic devices [J]. Scientometrics, 2010, 83(3): 765-781.

[5] Tu Y N, Seng J L. Indices of novelty for emerging topic detection[J]. Information Processing & Management, 2012, 48(2): 303-325.

[6] 殷蜀梅. 判断新兴研究趋势的技术框架研究[J]. 图书情报知识, 2008(3): 76-80.

[7] 刘玉仙, Rousseau R. 新出现趋势识别和分析方法引介[J]. 科学学研究, 2009, 27(7): 995-996.

[8] Leydesdorff L, Rafols I. A global map of science based on the ISI subject categories[J]. Journal of the American Society for Information Science and Technology, 2008, 60(2): 348-362.

[9] Boyack K W, Klavans R, Börner K. Mapping the backbone of science[J]. Scientometrics, 2005, 64(3): 351-374.

[10] Börner K, Chen C, Boyack K W. Visualizing knowledge domains [J]. Annual review of information science and technology, 2005, 37(1): 179-255.

[11] Chen C M. CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature [J]. Journal of the American Society for Information Science and Technology, 2005, 57(3): 359-377.

[12] Mane K K, Börner K. Mapping topics and topic bursts in PNAS [J]. Proceedings of the National Academy of Sciences of the United States of America, 2004, 101(Suppl 1): 5287-5290.

[13] 王晓光. 科学知识网络的形成与演化（Ⅰ）：共词网络方法的提出[J]. 情报学报，2009，28(4)：599-605.

[14] 王晓光. 科学知识网络的形成与演化（Ⅱ）：共词网络可视化与增长动力学[J]. 情报学报，2010，29(2)：314-322.

[15] Cobo M J, López-Herrera A G, Herrera-Viedma E., et al. An approach for detecting, quantifying, and visualizing the evolution of a research field: A practical application to the fuzzy sets theory field [J]. Journal of Informetrics, 2011,5(1): 146-166.

[16] Sun J, Faloutsos C, Papadimitriou S, et al. GraphScope: parameter-free mining of large time-evolving graphs, [C]//Proceedings of Knowledge Discovery in Databases:KDD. New York: ACM, 2007: 687-696.

[17] 钱铁云，李青，许承瑜. 面向科技主题发展分段的社区核心圈技术[J]. 计算机科学与探索，2010,4(2)：170-179.

[18] 吴斌，王柏，杨胜琦. 基于事件的社会网络演化分析框架[J]. 软件学报，2011,22(7):1488-1502.

[19] Ding Y. Community Detection: Topological vs. Topical [J]. Journal of Informetrics, 2011, 5(4): 498-514.

[20] Newman M E J. Detecting community structure in networks[J]. The European Physical Journal B, 2004, 38: 321-330

[21] Palla G, Derényi I, Farkas I, et al. Uncovering the overlapping community structure of complex networks in nature and society[J]. Nature, 2005(435): 814-818.

[22] Ball B, Karrer B, Newman M. Efficient and principled method for detecting communities in networks [J]. Physical Review E, 2011, 84(3): 36103.

[23] Lancichinetti A, Fortunato S. Community detection algorithms: A comparative analysis [J]. Physical Review E, 2009, 80(5): 56117.

[24] Blondel V D, Guillaume J-L, Lambiotte R, et al. Fast unfolding of community hierarchies in large networks [J]. Journal of Statistical Mechanics: Theory and Experiment, 2008, P1008: 1742-5468.

[25] McCain K W. Assessing an author's influence using time series historic graphic mapping: The oeuvre of Conrad Hal Waddington (1905-1975) [J]. Journal of the American Society for Information Science and Technology,2008,59(4): 510-525.

[26] Wallace M L, Gingras Y, Duhon R. A new approach for detecting scientific specialties from raw cocitation networks [J]. Journal of the American Society for Information Science and Technology, 2009, 60 (2): 240-246.

[27] Guimer'a R, Sales-Pardo M, N. Amaral L A. Classes of complex networks defined by role-to-role connectivity profiles[J]. Nature Phys, 2007 (3): 63-69.

[28] Palla G, Barabasi A L, Vicsek T. Quantifying social group evolution [J]. Nature, 2007, 446 (7136): 664-667.

[29] Berger-Wolf T Y, Saia J. A framework for analysis of dynamic social networks [C]// Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. New York: ACM, 2006: 523-528.

[30] Costa L F, Rodrigues F A, Travieso G, et al. Characterization of complex networks: A survey of measurements [J]. Advances in Physics, 2007, 56(1):167-242.

[31] Rosvall M, Bergstrom C T. Mapping change in large networks [J]. PloS one, 2010,5(1):e8694.

[32] Newman M E J. Fast algorithm for detecting community structure in networks [J]. Physical Review E, 2004, 69(6):66133.

[33] Newman M E J, Girvan M. Finding and evaluating community structure in networks [J]. Physical review E, 2004, 69(2):26113.

[34] C-DBLP. C-DBLP 主页，以作者为中心的学术搜索网站[EB/OL]. [2012-03-16]. http://www.cdblp.cn/.

（责任编辑　化柏林）