基于叙词表的关键词共现网络优化*

曹丽珠 宋培彦

（中国科学技术信息研究所 北京 100038）

〔摘要〕 以“白血病”为例提出基于叙词表优化关键词共现网络的方法，形成共现网络优化模型，采用模糊聚类方法对基于叙词表的数据进行优化聚类，指出该方法有助于优化聚拢共现网络，提高共现网络的中心聚集度和密度。

〔关键词〕 叙词表；模糊聚类；知识组织

〔中图分类号〕R－056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2018.05.014

Optimization of Keywords Co－occurrence Network Based on Thesaurus *CAO Li－zhu, SONG Pei－yan, Institute of Scientific and Technical Information of China, Beijing 100038, China*

〔**Abstract**〕 Taking "leukemia" as an example, the paper sets forward the method optimizing keywords co－occurrence network based on thesaurus to form the co－occurrence network optimization model, which carries out optimization clustering on data based on thesaurus through fuzzy clustering, and it points out that the method is conducive to optimization and clustering of the co－occurrence network and increase in the central concentration and density of the co－occurrence network.

〔**Keywords**〕 Thesaurus; Fuzzy clustering; Knowledge organization

1 引言

由关键词及其共现关系形成的网络被称为共现网络，它是以关键词作为知识单元构建的知识网络[1]。由于关键词具有模糊性、非结构化以及语义关系不明等缺点，在构建网络时容易造成知识网络庞大且过于发散的问题。而叙词表作为受控词表通常有明确的语义关系和严谨的词法规范，以规范化、具有明确概念含义的叙词为基本成分[2]，因此基于叙词表对共现网络进行优化有望解决该问题。本研究目标是通过利用叙词表的优点改进用户关键词的质量，实现基于叙词表的共现网络优化，同时对优化过的共现网络进行模糊聚类分析，更精细地揭示关键词节点之间的语义关系，对完善知识组织、提高知识服务效率将具有一定意义。

2 相关研究

构建关键词共现网络主要是进行共词分析，共词分析法即对词与词之间的共现关系进行分析，当前对共词分析方法的研究更多是侧重于以共词分析过程为主线，探究分析对象的改进、测度指标改进、可视化方法调整和融合其他方法等方面的理论研究，以及基于词、主题、时间维度和拓展应用4

〔修回日期〕 2018－03－21

〔作者简介〕 曹丽珠，硕士研究生；通讯作者：宋培彦，博士，副研究馆员。

〔基金项目〕 2016 国家社会科学基金项目“基于知识组织的科研项目评审专家发现研究”（项目编号：16BTQ079）。

个层次的具体应用研究[3]。相比单纯的词频统计方法，共词分析方法不仅注重关键词的文档频率，更加注重其相关性，从而能够更好地揭示关键词之间的语义关系[4]。国内外在共词分析研究大多是从研究学科领域结构的角度出发，对具体的学科领域进行实证分析。近几年开始对共词分析方法进行改进，如 Saason 等[5]提出基于共词分析方法扩展概念图的研究模型，使用网页计量网络计数来改进相似性度量，结论显示该方法可以延伸应用到其他领域。王玉林等[6]针对共词分析方法存在的共现词对的同量不同质问题、共词分析结果解释的不一致问题等，提出一种细粒度语义共词分析方法。冷伏海等[7]借鉴数字线划地图（Digital Line Graphic，DLG）关联挖掘算法提出基于位向量的三元共词分析算法和基于坐标图的三元共词结果分析方法，以主题图为指导的共词分析方法能够有效克服共词分析中的高频词孤立问题，社团主题更鲜明。上述方法为构建专业领域的术语词群提供有益的参考，在上述研究基础上以概念为核心、构建具有较高相关度的共现知识网络在方法上是可行的。同时由于关键词具有模糊性、非结构化以及语义关系弱的问题，需要严格规范的词表来约束，因此本文重点研究基于叙词表优化关键词，构建关键词共现网络，对关键词共现网络进行模糊聚类分析，为智能检索、个性化推荐等实际应用提供有效的知识基础。

3 关键词共现网络优化模型

3.1 整体优化模型

3.1.1 叙词表语义关系 根据国际标准化组织《ISO25964 信息与文献 - 叙词表与其他词表的互操作 - 第 2 部分：与其他词表互操作》，叙词表主要包括等同、等级和相关3种语义关系，主要通过这3 类关系来优化关键词网络。共现网络的等同、等级和相关关系归并过程，见图 1。图 1 - 1 所示共有 5 个节点，连线上的数值为两两之间的权重，节点 A 与节点 B、C 都相连，节点 C 与节点 D、E 相连。节点 C 与 D 从字面来看两者相似，经过查阅相关资料后发现 C 为规范词，D 为非规范词，即两者为同义关系，则删去节点 D 并将两者之间的权重加到 A 与 C 之间的权重上，见图 1 - 2。图 1 - 1 中节点 C 与 E 相连，查阅叙词表后发现 E 是 C 的上位词，则将 C 与 E 交换位置，见图 1 - 3。另外对于相关关系则需要通过计算词之间的语义相关度来确定。

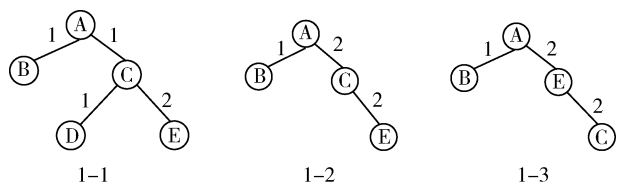

图 1 共现网络的等同、等级和相关关系归并过程

3.1.2 《中文医学主题词表》（Chinese Medical Subjects Headings，CMeSH） 中国医学科学院医学信息研究所出版的《医学主题词表》中文版，用于中文医学文献的标引、编目和检索。医学主题词为同一概念具有不同表达方式的词语提供规范标准的用语，使文献加工处理达到高度的统一，为文献查询提供便利。

3.1.3 基于叙词表的关键词共现网络优化模型（图 2） 主要包括数据层和语义层。在数据层进行数据收集、清洗后构建关键词共现网络，作为后续处理的基础。在语义层依据叙词表的 3 种语义关系进行共现网络优化，分别进行归并处理，提高知识网络的密度和关联性。对共现网络进行密度和中心势分析，用于改进关联效果。

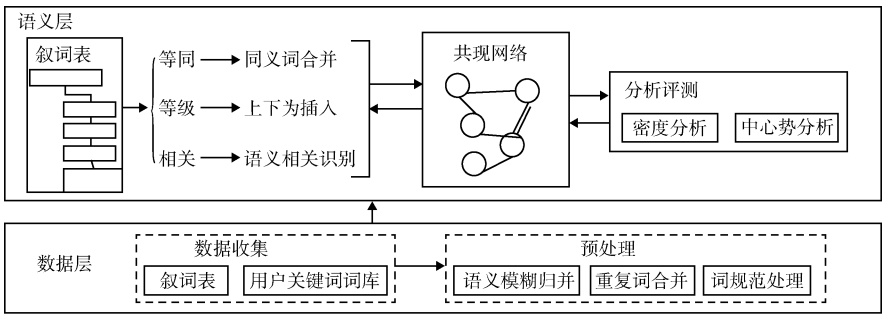

图2 基于叙词表的关键词共现网络优化模型

3.2 模糊聚类方法

3.2.1 概述 社群是社会网络结构中具有内聚性的特定团体[8]，同一社群的各节点间趋于内部联系紧密、外部联系稀疏[9-10]。从社群角度看可以将这些术语概念分割成几个不同的群体，不仅可以深入洞悉社群内部结构，而且有助于从群体视角理解整个网络的结构和功能，揭示各群体之间的关系[11]。模糊聚类算法是社群划分的一种方法，而同时聚类是一个无监督学习过程[12]，在聚类前很难根据经验知识确定聚类数，因此本研究选用模糊聚类算法对关键词进行聚类分析，动态挖掘出其中的深度关系。

3.2.2 数据标准化 由于样本的量纲和数量级不一定相同，故在运算过程中可能突出某数量级特别大的特性指标对分类的作用，而降低甚至排除某些数量级很小的特性指标的作用，数据规格化使每个指标值统一于某种共同的数值特性范围。

3.2.3 构造模糊相似矩阵 聚类是按照某种标准来鉴别元素间的接近程度，将彼此接近的对象归为一类，一般利用夹角余弦、Pearson 或 Jaccard 来计算相似矩阵。

3.2.4 模糊聚类 通过以上两个步骤建立起来的模糊关系一般只具有自反性和对称性，不满足传递性，所以利用其构造一个新的模糊等价矩阵，然后依次设定截集 λ 进行动态聚类。

4 实证研究

4.1 数据来源

以《中国图书馆分类法》中 R733.7“白血病”为依据，从万方数据库中抽取部分数据形成 268 × 268 的共现矩阵，见表 1。使用万方医学网的 MeSH 主题词获取“白血病”部分主题词，发现白血病共分为 3 大类，其中“白血病，实验性”对应的主题词有 5 个，“白血病，淋巴样”对应的主题词有 15 个，“白血病，髓样”对应的主题词有 21 个。

表1 “白血病”共现矩阵

主题词	白血病	白血病，淋巴样	白血病，髓细胞，急性	白血病，髓样，急性	…
白血病	0	0	0	0	…
白血病，淋巴样	0	0	0	0	…
白血病，髓细胞，急性	0	0	0	0	…
白血病，髓样，急性	0	0	0	0	…
…	…	…	…	…	…

4.2 关键词处理

4.2.1 概述 虽然研究人员在撰写论文时会尽量使用规范词进行标引，但仍不排除会根据个人对某些知识的理解给出一些非规范标引词。本研究根据叙词表——《医学主题词表》中的相关部分对这268 条数据进行规范化、专业化处理，制定几条准则，为处理方便本研究所有关键词都对应唯一ID 号。

4.2.2 关键词规范化处理 （1）处理符号、语义模糊的关键词。如“白血病，慢性，髓性”在叙词表中并没有找到这一分类，通过网络搜索“慢性髓性白血病”发现其属于髓样这一部分，为方便识别这类词将其格式改为“某某白血病”类型。另外像“白血病急”这一类词可能因为书写错误或抽取过程中出现的误差使词不完整，通过查看该词的频次发现仅为1，影响甚微所以删除。在作者编写论文或在抽取过程中导致一些符号或英文格式不规范，对这类词进行逐一处理。如“白血病：U937 细胞？”中通过查阅相关资料发现冒号和问号无意义，因此删除这类无意义的符号。疾病类关键词经常会出现英文，统一将其转换为英文半角大写，另外对于“-”，查阅相关医学资料确定是否需要含有该字符串。（2）合并与转化重复关键词。在抽取过程中可能因标点符号不一致会使同样的词收录两遍，通过人工检查，一一对比，进行合并删除处理。在作者编写论文时会根据内容或个人理解对一些词进行前后对换，如“白血病，急性，早幼粒细胞”和“白血病，早幼粒细胞，急性”都是对急性早幼粒细胞白血病的描述，因此对这类词进行合并去重。在医学上有些词虽然说法不同但指向对象一致，如“白血病，急性，髓细胞性”和“白血病，急性髓细胞”都是指急性髓细胞白血病，这类词数量有限，通过人工搜索并查阅相关资料进行对比，进行合并去重。

4.2.3 游离散点组优化处理（图3） 经过规范化处理后发现网络周围有部分单独小组，距离中心较远，所以对这类词进行处理。图3 中“白血病，T 细胞，慢性”、“白血病大颗粒淋巴细胞”、“白血病，单核细胞，急性/诊断”和“白血病，单核细胞，急性/并发症”等节点游离在周围且与其他节点没有关联。（1）处理等同关系的关键词。“白血病，T 细胞性，大颗粒淋巴细胞性”和“白血病大颗粒淋巴细胞”均是指白血病大颗粒淋巴细胞，所以删除“白血病，T 细胞性，大颗粒淋巴细胞性”并对共现矩阵中的“白血病大颗粒淋巴细胞”增加相应权重。（2）处理等级关系的关键词。“白血病，T 细胞，慢性”与“白血病，T 细胞，急性”均属于“白血病，T 细胞”且两者为等级关系，因此将两者相连，在共现矩阵中增加权重。同样通过查阅叙词表，“白血病大颗粒淋巴细胞”也属于“白血病，T 细胞”且与上述两个关键词为等级关系，因此处理方式相同。通过上述处理目前得到关键词数237 个。

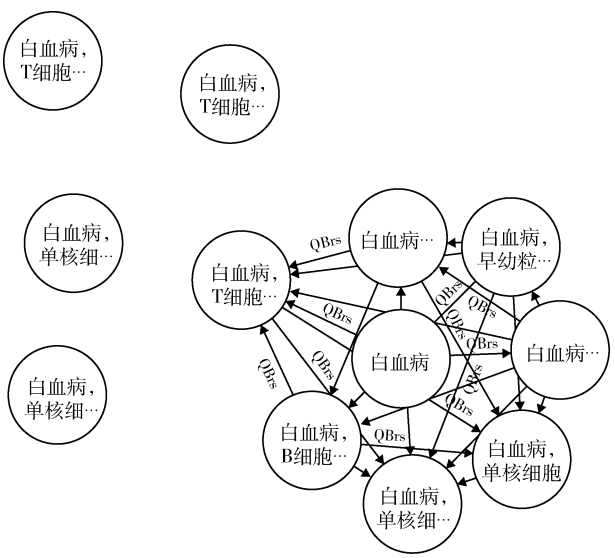

图3 游离散点组优化处理

4.3 模糊聚类

4.3.1 数据相似及标准化处理 将上文处理过的白血病关键词共现矩阵导入SPSS Statistics21，进行分析-相关-距离处理，选择Pearson 进行变量间相似性计算，得到相似矩阵。过滤后的“白血病”相似矩阵，见表2。

表2 过滤后的“白血病”相似矩阵

主题词	白血病	白血病，淋巴样	白血病，髓细胞，急性	白血病，髓样，急性	白血病，早幼粒细胞，急性	白血病，急性	...
白血病	1	0	0	0.114	0.027	0.021	...
白血病，淋巴样	0	1	0	0	0	0	...
白血病，髓细胞，急性	0	0	1	0	0	0	...
白血病，髓样，急性	0.114	0	0	1	0.019	0.15	...
白血病，早幼粒细胞，急性	0.027	0	0	0.019	1	0.017	...
白血病，急性	0.021	0	0	0.015	0.017	1	...
...	...	...	...	...	...	...	...

4.3.2 构造模糊等价矩阵 将相似性矩阵导入MATLAB，通过设定一个值作为找到传递闭包的标志，利用矩阵自乘得到模糊等价矩阵 N，因篇幅有限只截取一部分数据。

$$N = \begin{bmatrix} 1 & 0 & 0 & 0.263 & \dots \\ 0 & 1 & 0 & 0 & \dots \\ 0 & 0 & 1 & 0 & \dots \\ 0.263 & 0 & 0 & 1 & \dots \\ \dots & \dots & \dots & \dots & \dots \end{bmatrix}$$

4.3.3 聚类及结果展示 通过利用阈值转换模糊等价矩阵中的值，若≥阈值则转换为 1，若 < 阈值则转为 0，最后根据比较行向量得到聚类结果。在模糊等价矩阵 N 中可取阈值：0.099；0.012；0.106；0.125；0.192；0.263；0.296；0.313；0.404；0.439；0.497；0.575；0.601；0.706；0.961，得到相应的聚类矩阵 Q，经整理最终得到以下聚类结果，每个关键词都有唯一 ID。不同阈值对应的聚类结果（部分），见表 3。

表3 不同阈值对应的聚类结果（部分）

阈值	聚类结果
0.012	{1，4，5，6，11，15，19，20，26，29，34，52，55，58，59，60，61，62，65，73，75，76，81，83，85，103，104，107，121，123，129，132，139，140，143，147，149，151，152，153，154，156，162，169，170，172，173，175，185，186，187，189，194，197，204，209，211，213，214，217，218，230，232，233}，其余关键词各自为一类
0.099	{1，4，5，6，11，15，19，20，26，34，52，55，65，75，76，103，104，107，121，129，131，140，154，156，162，169，170，172，173，175，185，186，187，189，194，197，204，209，211，213，214，217，218，230，232，233}，其余关键词各自为一类
0.106	{1，4，5，6，11，15，19，20，26，34，55，65，75，76，103，104，107，121，129，131，140，154，156，162，169，170，172，173，175，185，186，187，189，194，197，204，209，211，213，214，217，218，230，232，233}，其余关键词各自为一类
0.125	{1，4，5，6，11，20，26，34，52，55，65，75，76，103，104，107，121，129，131，140，154，156，162，169，172，173，175，185，186，187，189，194，197，204，209，211，213，214，217，218，230，232，233}，其余关键词各自为一类
0.192	{1，4，5，6，11，20，26，34，52，55，65，75，76，103，104，107，121，129，131，140，154，156，162，169，172，173，175，185，186，187，189，194，197，204，209，211，213，214，218，230，232，233}，其余关键词各自为一类
0.263	{1，4，5，6，11，20，26，34，52，55，65，75，76，103，104，107，121，129，131，140，156，162，169，172，173，175，185，186，187，189，194，197，204，209，211，213，214，218，230，232，233}，其余关键词各自为一类

4.3.4 二次聚类结果 通过比较这 15 种聚类结果，最终得出阈值为 0.263 为最优结果，即其中 41 个关键词为一类，其他 196 个关键词各自为一类，利用 JAVA 编程将相似矩阵变成关系 3 元组（关键词－关系－关键词），导入 NEO4J 数据库。通过观察 NEO4J 示意图，发现存在很多漂浮在周围且无关联的节点，对比关键词频次表，其余的 196 个关键词频次几乎都是 1，且基本没有关系相联，对结果的影响甚微，故删去这些节点，然后基于 CMeSH 叙词表只对这 41 个关键词对进行二次聚类。将聚类结果再次导入 NEO4J 数据库，模糊聚类可视化结果，见图 4。

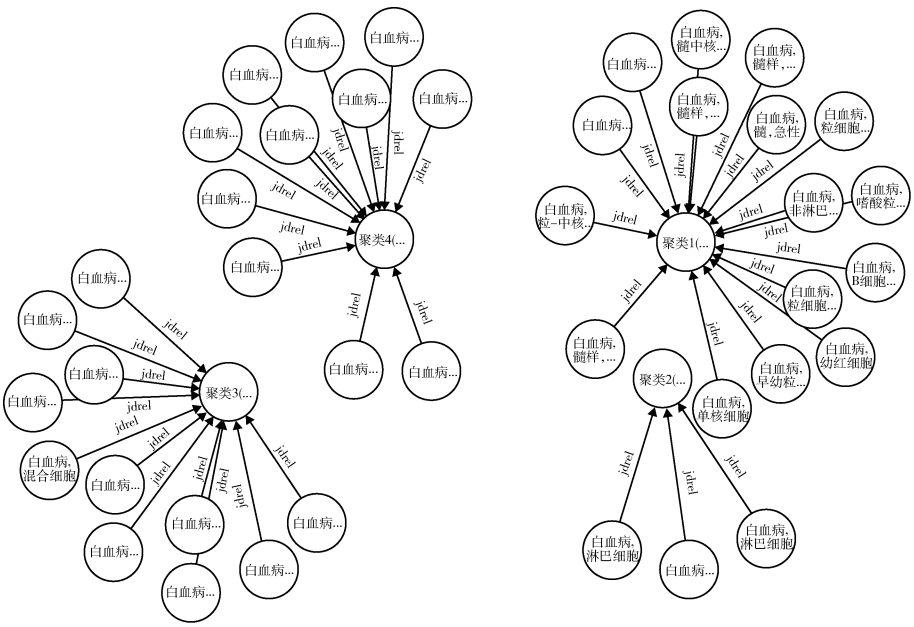

图 4 模糊聚类可视化结果

4.3.5 聚类结果分析 从 CMeSH 主题词中可以发现白血病共被分为 5 大类，其中髓样和淋巴样类目下所含主题词最多，这与聚类结果一致。在 NEO4J 中使用 Cypher 查询可以得到每个节点及其与之相关节点的关系图，在“白血病，髓样”中与“白血病，髓样，急性”相关的节点有 63 个。另外“白血病，早幼粒细胞，急性”，“白血病，单核细胞”，“白血病，单核细胞，急性”和“白血病，非淋巴细胞，急性”相关节点分别为 62、58、57 和 56 个，位于前列。所以可初步确定在“白血病，髓样”中目前研究讨论最多的是这几种。同样在“白血病，淋巴样”中“白血病，B 细胞，侵袭性”，“白血病，T 细胞，急性”，“白血病，淋巴细胞，急性/护理”和“白血病，淋巴细胞，急性/免疫学”相关节点为 60、59、44 和 43 个，可知这 4 个节点在该聚类中关联性最强，另外“白血病，细胞”中“白血病，混合细胞”相关节点最多，为 54 个并与其他节点有较大差距，在白血病中混合细胞是指由髓细胞和淋巴细胞共同累及的细胞，而这两类细胞正是白血病中占有率最高的，因此与结果基本一致。在“白血病，其他”部分各个节点之间差距不大，基本可以确认主要集中在白血病病理学特征和护理方面。

4.4 社会网络分析

运用定量分析的方法测量网络结构，刻画网络的具体形态和特性。鉴于此本研究利用社会网络分析方法对共现网络结构和特征进行分析。网络基本特征包括密度、中心性分析等，用以描述整个网络的规模和紧凑程度。网络密度可用于刻画网络中节点间相互连边的密集程度，定义为网络中实际存在的边数与可容纳的边数上限的比值。一个具有 N 个节点和 L 条实际连边的网络，其网络密度公式为：$d(G) = \frac{2L}{N(N-1)}$。中心势是指比较网络的边缘点

和中心点的中心度情况，如果一个网络很集中，那么中心点的中心度高而边缘点中心度低；如果一个网络很稀疏，那么中心点、边缘点的中心度差异较小，因此网络中心势衡量整个网络向中心聚集的程度，公式为 $C_{AD} = \frac{\sum_i (C_{ADmax} - C_{ADi})}{(n-1)(n-2)}$，$C_{ADmax}$指网络中节点中心度的最大值，$C_{ADi}$指网络中第 i 个节点的中心度。将处理前后的白血病数据分别导入 UCINET 进行密度分析和网络中心势分析，网络特征分析结果，见表4。从表中数据可知处理前后的网络密度都较低，因为网络中存在一些无关联的节点但这些节点与白血病又相关所以并没有做删除处理，但网络密度提高了 17.2%，网络更加紧密，这可以说明达到了一定优化效果。网络中心势也由 0.96% 提高到 1.08%，说明优化过的共现网络更加集中在网络中影响力大的节点。

表4 网络特征分析结果

指标	处理前	处理后
密度	0.0024	0.0029
网络中心势（%）	0.96	1.08

5 结语

本研究以“白血病”为例，提出基于叙词表优化关键词构建共现网络的模型与方法，使用模糊聚类算法进行聚类分析。研究结果表明医学主题词表严谨、规范的类目层级关系与文献关键词的全面、动态相结合，能优化关键词，使构建的共现网络更加清晰直观。本研究为保证关键词的全面性，保留了低频词，这可能导致共现网络比较发散，今后通过聚类等方法提高语义关联性仍需要进一步研究。同时叙词表一般规模较小，对开放领域的共现网络的优化作用还需要完善，未来可以通过与本体、术语库等其他知识资源相结合，探索其在不同领域的适用性，形成面向计算机自动处理的知识组织方法，促进知识组织的精准化和自动化。

参考文献

1 王众托. 知识系统工程（第2版）[M]. 北京：科学出版社，2016：207-213.

2 贾君枝，孙智超，郜杨芳. 基于受控词表的医学资源社会化标签推荐研究 [J]. 情报学报，2013，32（12）：1326-1332.

3 唐果媛，张薇. 国内外共词分析法研究的发展与分析 [J]. 图书情报工作，2014，58（22）：138-145.

4 杨建林. 关键词选择策略及其对共词分析的影响 [J]. 情报学报，2014，33（10）：1083-1090.

5 Saason, Ravid, Nava, et al. Improving Similarity Measures of Relatedness Proximity: Toward Augmented Concept Maps [J]. Journal of informetrics, 2015, (3): 1751-1577.

6 王玉林，王忠义. 细粒度语义共词分析方法研究 [J]. 图书情报工作，2014，58（21）：73-80.

7 冷伏海，王林，李勇，等. 基于文献关键词的三元共词分析方法——以知识发现领域为例 [J]. 情报学报，2011，30（10）：1072-1077.

8 Tang Lei, Liu Huang. 社会计算：社区发现和社会媒体挖掘 [M]. 北京：机械工业出版社，2013：21，29-30.

9 Wilkinson D M, Huberman B A. A Method for Finding Communities of Related Genes [J]. Proceedings of the National Academy of Sciences, 2004 (suppl 1): 5241-5248.

10 Radicchi F, Castellano C, Cecconi F, et al. Defining and Identifying Communities in Networks [J]. Proceedings of the National Academy of Sciences of the United States of America, 2003, (9): 2658-2663.

11 李纲，任佳佳，毛进，等. 专利权人合作网络的社群结构分析——以燃料电池电动汽车专利为例 [J]. 情报学报，2014，33（3）：267-276.

12 B. W. Kernighan, S. Lin. A Efficient Heuristic Procedure for Partitioning Graphs [J]. Bell Systems Technical Journal, 1970, (2): 291-307.