基于复杂网络的知识组织系统概念社区发现*

殷希红，乔晓东，张运良
（中国科学技术信息研究所，北京 100038）

摘要：将复杂网络的理论引入到知识组织系统的表示中，抽取知识组织系统中的概念及概念间的关系，构建复杂网络。利用Walktrap社区发现算法，发现复杂网络中的概念社区，以助于用户输入种子概念时，仅返回对应的社区。利用种子概念返回社区的准确性对社区发现的结果进行评价，论证该方法的有效性。本研究将以中国科学技术信息研究所已有的新能源汽车领域汉语科技词系统为例进行社区发现研究，发现该方法快速有效。

关键词：复杂网络；知识组织系统；概念社区；社区发现
中图分类号：G254.2；TP391
DOI：10.3772/j.issn.1673—2286.2014.08.007

1 引言

在叙词表等知识组织系统构建的过程中，都需要利用既有的资源。一般来说，无法将已有知识组织系统全部纳入新的知识组织系统中，而仅需要选择一部分，理想的情况是根据一部分种子概念选择出适当数量的相关概念，并保留这些关系，作为新的知识组织系统的一部分。但是抽取过程中很容易通过关系不断扩展得到非常大的一个集合，甚至是知识组织系统的全部，无法达到抽取适当数量相关概念的要求。

复杂网络是复杂系统的抽象，网络中的节点是复杂系统中的个体，节点之间的边则是系统中个体之间按照某种规则而自然形成或人为构造的一种关系[1]。Newman于2002年提出了复杂网络的社团结构概念[2]，杨格兰[3]认为网络社团结构指网络中的顶点可以分成组（块），组（块）内顶点之间的连接比较稠密，组（块）间顶点的连接比较稀疏。在大型的复杂网络中进行社区发现具有重要的研究意义和实用价值。社会网络[4]中的社区代表根据兴趣或背景而形成的真实的社会团体；引文网络中的社区代表针对同一主题的相关论文；万维网中的社区就是讨论相关主题的若干网站[5]；而生物化学网络或者电子电路网络中的社区可以是某一类功能单元[6]。复杂网络的社区发现算法就是发现不同类型的复杂网络中固有的社区结构，揭示复杂网络中隐含的关系及规律，以助人们清晰地看出复杂网络的内在结构，预测复杂网络的行为。

本文拟将复杂网络理论引入到知识组织系统的表示当中，把知识组织系统中的每一概念抽象为复杂网络中的节点，概念间的关系抽象为节点间的边。利用复杂网络社区发现的已有算法，将知识组织系统构建的复杂网络划分为不同的社区，每个社区代表着针对某一主题概念的相关术语。这样在用户输入种子概念的时候，就能够返回有关的一个或几个社区，从而支持利用既有资源完成相关领域的新的知识组织系统构建。本研究将以中国科学技术信息研究所已有的新能源汽车领域汉语科技词系统[7]为例，进行社区发现研究，考察概念社区发现的效果。

2 复杂网络社区发现算法

复杂网络的社区发现研究已有近十多年的时间，有不少经典的社区发现算法被提出，1970年，Kernighan和Lin提出一种试探优化法划分网络中的社区结构，简称K-L算法[8]。它是一种基于贪婪算法原理

* 本研究得到国家自然科学基金项目“面向特定情报分析应用的知识组织系统快速构建关键问题研究”（编号：71203208）、国家“十二五”科技支撑计划课题“面向外文科技文献信息的超级科技词表和本体建设”（编号：2011BAH10B01）和中国科学技术信息研究所重点工作项目“汉语科技词系统建设与应用工程”（编号：ZD2012-3-2）资助。

将网络划分为两个大小已知的社区的二分法，将一个网络节点图分割成两个相等的节点集合，使连接两个社区的边权最小。其基本思想就是为网络的划分引进一个增益函数Q，定义为两个社团内部的边数减去连接两个社团之间的边数，然后寻找使Q值最大的划分方法。该方法对分别属于两个社区的每个节点对，计算如果交换这两个节点可能得到的Q的增益$\Delta Q=Q_{交换后}-Q_{交换前}$，然后交换最大的ΔQ对应的节点对，同时记录交换以后的Q值。规定每个节点只能交换一次。重复这个交换过程，直到某个社团内所有的节点都被交换一次为止[9]。但是该方法仅能将网络划分为两个社区，有一定的局限性，不适合本实验中的概念社区划分。2002年Girvan和Newman提出了基于边介数的分裂算法，即GN算法[10]。该算法采用的启发式规则为：社区间链接的边介数（Edge Betweenness）应大于社区内链接的边介数，其中每个链接的边介数被定义为“网络中经过该链接的任意两点间最短路径的条数”[11]。该算法是每计算一次然后移除复杂网络中边介数最大的链接，直到将网络划分为一个个独立的社区为止，其时间复杂度为O（m^2n），其中m表示链接数，n表示结点数。该算法的缺点是计算速度较慢，是由于边介数计算开销过大引起，因此该算法只适合处理中小规模网络（其包含的结点数通常<10^3）。基于GN算法，Tyler[12]引入了统计方法，提出近似的GN算法：采用蒙特卡洛方法估算出部分链接的近似边介数，而不去计算全部链接的精确边介数。而这种方法计算速度的提高是以牺牲聚类精度为代价的。而对于大规模复杂网络社区发现的处理，Pascal Pons与Matthieu Latapy[13]提出了基于随机游走划分社区结构的Walktrap算法，该方法针对无向网络图G=(V, E)，假设G为连通图，并且n=|v|表示节点的个数，m=|E|表示边的数量。其初始条件为每个结点为一个单独的社区，然后逐步合并可使结点和它所在社区之间的平方距离均值达到最小的两个社区，每一步都要更新社区之间的距离，其中两个结点之间的距离对应于它们的相似度，即在一个离散的随机游走过程中，它们之间的方向转移概率[14]。该算法具有三点优势：第一，该算法可以较好地获取复杂网络中的社区结构。第二，该算法可以快速有效地实现社区划分，在最坏的情况下其时间复杂度为$O(mn^2)$，空间复杂度为$O(n^2)$。而在大多数现实网络中其运行时间复杂度为$O(n^2\log n)$，空间复杂度为$O(n^2)$。第三，使用凝聚的算法来有效地计算网络中的社区结构。总之，相对其他社区发现算法，Walktrap算法可以取得较好的社区划分结构和较快的运行效率。因此，本文采用Walktrap方法用于复杂网络概念社区的发现。另外该方法利用Newman等[15,16]提出的模块度Q来衡量最好的社区划分结果，模块度Q的物理意义是：网络中连接社团内部的两个节点的边占网络总的边数的比例，减去在同样的社团结构下任意连接这两点的边的比例的期望值。Q值越大说明网络的社团结构越明显。

3 方法

3.1 抽取概念及概念间关系

从新能源汽车知识组织系统中抽取出概念及概念间的关系，将每一个概念及与其有关系的概念均抽取出来存储到MySQL数据库中，存储方式如表1所示。

表1 概念及概念间的关系

termA_id	termA	relation_id	title	termB_id	termB
0	生物燃料	15	子类	9717	第一代生物燃料
0	生物燃料	14	类属	9715	可再生燃料
0	生物燃料	14	类属	329	可再生能源
0	生物燃料	14	类属	7401	清洁汽车燃料
0	生物燃料	62	输入-设备	9519	生物燃料汽车
0	生物燃料	15	子类	10043	第三代生物燃料
0	生物燃料	15	子类	9724	第二代生物燃料
0	生物燃料	15	子类	9744	第1.5代生物燃料
0	生物燃料	20	异名同义	10274	生物质燃料

表1 概念及概念间的关系（续）

termA_id	termA	relation_id	title	termB_id	termB
0	生物燃料	14	类属	3832	汽车代用燃料
0	生物燃料	68	产品-过程	2176	生物质气化技术
0	生物燃料	68	产品-过程	9748	生物质高温裂解
0	生物燃料	68	产品-过程	9749	乳化法
0	生物燃料	68	产品-过程	9750	水蒸气重整

其中termA和termA_id为概念A及概念A对应的编号，而title和relation_id为termA和termB的关系类型及关系对应的编号，termB和termB_id为概念B及概念B对应的编号。

3.2 构建复杂网络

根据知识组织系统中的概念及概念间关系构建复杂网络，知识组织系统中的每一个概念在网络中均有一个节点与之对应，而概念间的关系构成节点间的连边，这样便形成一个无向图。若输入时未对边作加权处理，系统会默认边的权值为1，但也可以对边作加权处理，权值可根据自身需要设置。若两个概念间存在多种关系，即两个节点间存在多条边，而在这里多条边将被视为一条边，其边的权值则为这两节点间所有边的权值之和。在实验中分别采用对边作未加权和加权处理的两种复杂网络进行社区发现。

3.3 社区发现

采用Walktrap方法[13]在复杂网络中发现概念社区。该方法采用在图中随机游走的方式，认为游走会“陷入”到互相连接密集的部分，即为社区。基于Walktrap划分复杂网络概念社区的步骤如下：

（1）将知识组织系统构成的复杂网络中的每个节点看成一个个独立的概念社区。

（2）计算所有邻接节点之间的距离。其对应于节点之间的相似度，为一节点经过t步到达另一节点的方向转移概率，该计算基于以下假设：若两节点在同一社区，则这两个节点到达网络中的任一节点的概率是基本相等的。

（3）合并可使节点和它所在社区之间的平方距离均值达到最小的两个概念社区，成为一个概念社区。

（4）更新社区之间的距离（实际上只计算了邻接社区之间的距离），其计算方式与节点间距离计算方式相似。

（5）计算衡量社区划分质量的模块度值Q。

（6）重复步骤（3）、（4）、（5），得到对应不同社区划分的模块度Q。

（7）取最大的模块度值对应的社区划分作为最后的概念社区划分结果。

本实验的运行环境为Windows XP操作系统，2.00GHz的CPU，2G的内存，使用软件为walktrap.exe。使用walktrap [input_file] [-o output_file] [-i index_file] [options]命令进行社区划分，其中“input_file”为存储复杂网络的文件，“output_file”为保存输出结果的文件。而“-i index_file”为索引文件，“options”为该算法的各种参数设置，使用“-tx -dl -s -b”，其中“-tx”中的“x”设置随机游走的长度。“-dl -s -b”仅输出该参数设置下最大模块度对应的社区划分结果并保存在“output_file”中，输出结果中包括最大模块度的值Q及最大模块度对应的社区划分，并且记录在该实验环境下的实验时间。

4 实验

4.1 数据来源

选取新能源汽车领域汉语科技词系统领域的12294个概念作为语料，概念间语义关系已确立了15个一级关系，79个二级关系，这些概念间的关系数量共有57101个。平均每个概念包含的关系数量约为5。概念间关系类型及数量如表2所示。

从表1可以看出概念间的层级关系和组成关系占大部分，其中层级关系中的类属关系和子类关系，组成关系中的隶属于和拆解为占关系的多数。

表2 新能源汽车词系统关系空间表

一级关系类型	二级关系类型	数量	一级关系类型	二级关系类型	数量	一级关系类型	二级关系类型	数量
适用情况	用于	621	转变关系	替代	4	过程相关	过程-产品	109
	适用	124		转变（相继）	7		产品-过程	146
	受限	40		继承	9		过程-工具	15
触发条件	是条件	14	空间关系	相连接	113		工具-过程	5
	依据	21		相平行	2	对应关系	类比	890
	是前提	14		固定于	39		相似	74
	处境	11		来源	10		可替代	7
层级关系	类属	10432	因果关系	结果（因-果）	32		配合	136
	子类	11511		成因（果-因）	51		无关	3
	被类分依据	567		影响（部分因果）	232		现象-结论	2
等同关系	全称-缩略同义	247		受影响（有关）	254		结论-现象	1
	异名同义	1611		取决	23		主体-附件	54
组成关系	隶属于	12064	输入输出	驱动	16		附件-主体	48
	拆解为	12978		传递	14		概念-实例	105
控制关系	控制（操纵）	42		传送	1		实例-概念	65
	控制关系	1		材料-成品	665		表示（表征）	13
	调节	38		材料-设备	221		表明（反映）	10
	筛选	2		消耗品-设备	154		性能-主体	31
	避免（防止）	3		成品-材料	676		主体-性能	31
	限制（约束）	7		设备-材料	220		指标-主体	797
	取消	1		设备-输入	160		主体-指标	796
借助	借助	159	物理量相关	物理量-单位	23		过程-主体	11
	利用	150		单位-物理量	14		主体-过程	11
时间关系	前	2		物理量-度量工具	48		主体-计算	8
	后	2		度量工具-物理量	49		计算-主体	8
	同一时段	1		物理量-度量方法	8		可细分特性	16
				度量方法-物理量	1			

4.2 构建复杂网络

根据概念间关系分别构建未加权和加权复杂网络。对于构建概念的未加权复杂网络只要是概念间存在关系即构成了概念间的一条连边，而对于构建概念的加权复杂网络则需要对概念间关系进行加权处理，根据概念间关系的强弱将概念关系赋予不同的权值，将等同关系的权值设为1，层级关系的权值设为0.75，组成关系的权值设为0.5，而其他关系的权值统一设为0.25，表3中的weight列是为两个概念间关系赋予的权值。

新能源汽车汉语科技词系统中共选取出的概念数目为12994，而概念间的关系数量为57101，将概念抽象为节点，概念间的关系抽象为节点间的连边，因此构建的概念复杂网络中包括节点数为12994，而节点间构成的连边数为57101。

4.3 社区发现

根据walktrap要求的输入格式，如果概念间存在关系，则对于未加权复杂网络仅输入概念对应的编号，

表3 加权后的概念关系对应表

termA_id	termA	relation_id	title	termB_id	termB	weight
0	生物燃料	15	子类	9717	第一代生物燃料	0.75
0	生物燃料	14	类属	9715	可再生燃料	0.75
0	生物燃料	14	类属	329	可再生能源	0.75
0	生物燃料	14	类属	7401	清洁汽车燃料	0.75
0	生物燃料	62	输入-设备	9519	生物燃料汽车	0.25
0	生物燃料	15	子类	10043	第三代生物燃料	0.75
0	生物燃料	15	子类	9724	第二代生物燃料	0.75
0	生物燃料	15	子类	9744	第1.5代生物燃料	0.75
0	生物燃料	20	异名同义	10274	生物质燃料	1.00
0	生物燃料	14	类属	3832	汽车代用燃料	0.75

表4 复杂网络的社区划分数量和模块度值Q

	未加权复杂网络			加权复杂网络		
t	communities	模块度值Q	时间（s）	communities	模块度值Q	时间（s）
4	518	0.754482	7.3	540	0.768502	8.2
5	470	0.761112	12.6	481	0.781587	13.2
6	422	0.766952	19.7	415	0.788403	20.1
7	369	0.773101	26.1	380	0.791932	27.2
8	350	0.773763	32.8	342	0.79413	33.1
9	329	0.774323	39	329	0.795229	39.4
10	310	0.775323	44	316	0.799605	44.7

如（0　9717），说明节点0和9717间存在一条边，其默认权值为1；而加权复杂网络还需要在对应编号后面加上对应的权值，如（0　9717　0.75），说明节点0和9717间边的权值为0.75。然后利用社区划分命令，针对不同的游走长度t值进行复杂网络的社区划分实验，如表4所示为随游走长度t值的变化，未加权复杂网络和加权复杂网络的社区划分数量、模块度值Q的变化情况及实验所需时间。

从表4可以看出，随着t值的增大，社区划分的数量越来越少，而模块度值Q越来越大，所使用的实验时间越来越多。预测当t值增加到一定限度时，社区划分数量会趋于一个极限值，为最大连通子图的数量，而且模块度值也为最大值。但是该情况下，当输入种子概念的时候可能会导致大部分知识组织系统甚至整个知识组织的抽取，不符合该实验需求。因此t值的选择可在社区数量和模块度之间取一个折中，或者根据自身需求进行调整。

从表4还可以看出，本次实验所用时间相当少，以秒计算。在t值相同的情况下，加权复杂网络的模块度普遍高于未加权复杂网络的模块度，而加权复杂网络与未加权复杂网络所用的实验时间相差甚微，说明加权复杂网络的社区划分质量高于未加权复杂网络的社区划分质量。

5 结果评价

利用种子概念对复杂网络的社区发现结果进行检测，从“新能源词系统”中选取种子概念进行测试，将

表5 社区发现结果及其提取词条和关系的数量

种子词	所在社区	社区词条	社区关系数量	准确率
太阳能电池	community 23181	143	3173	35.7%
沼气池	community 23862	6	20	66.7%
醛	community 23000	10	26	90%
风能	community 23651	440	3767	59.8%
总计	——	599	10179	——

利用种子概念抽取出的社区中包含的词条与“新能源词系统”中与种子概念有关系的词条进行比较，得到社区发现结果的准确率。其计算如公式（1）所示：

$$\text{Precision}=\frac{\text{common } terms<kos1,kos2>}{\text{number } terms \text{ of kos1}} \quad (1)$$

其中common terms<kos1，kos2>为两知识组织系统中共同的词条数量，而number terms of kos1为利用种子概念在“新能源汽车词系统”中抽取的词条数量。在“新能源词系统”中提取“太阳能电池”、“沼气池”、“醛”、“风能”四条具有代表性的词条作为种子概念，在新能源汉语科技词系统中进行检索，以游走长度为4时的加权复杂网络社区发现结果为例，找到对应社区，并提取社区内部词条之间的关系，并利用准确率对社区发现结果进行评价，如表5所示。

从表4可以看出，社区发现结果大体可以达到较高的准确率，但是以“太阳能电池”作为种子词时，准确率略低，仅有35.7%。其一，是因为不同的词系统在收词的过程中侧重点不同，“新能源汽车词系统”中更倾向于汽车方面的词条，而“新能源词系统”中倾向与能源有关的词条。其二，是因为从“新能源汽车词系统”中抽出的词条数量过多，而“新能源词系统”中仅含有少部分词，这样即使抽出的相同词条数量多，但受准确率计算公式的影响，同样会造成准确率较低。但是由于概念社区发现的目的不是为了获取完整的一个新的词系统，而是得到一个构建新的知识组织系统可用的框架，因此不完全受准确率的影响。综上，说明该方法是可行的，可为构建知识组织系统提供概念及概念间的关系。并且种子概念所提取的社区中存在与种子概念有关系但尚未构建的概念，可抽取此类概念及概念间的关系丰富原有知识组织系统。

6 结语

本研究利用复杂网络的社区发现算法解决知识组织系统中根据种子概念仅抽取部分相关概念的问题。实验结果表明，本次实验可以达到较高的模块度值，社区划分质量较高，可以实现当用户输入某个种子概念时，仅返回对应的社区，并且该方法用时较少，说明该方法快速有效可行。但是该方法还存在以下不足：第一，在选取的新能源汽车汉语科技词系统中，理论上两个词之间如果有关联，应该就是两个关系，但是构建的过程中有疏漏，部分对等关系没有构建，这样便造成两节点间权值的缺失，影响社区划分算法的计算，对实验结果造成影响。第二，不能实现重叠社区的发现，在该方法中一个概念仅有一个社区相对应，而实际情况不一定如此。第三，该方法不能实现有向复杂网络的社区发现。因此，在之后研究中还需要融入其他聚类算法在重叠社区发现和有向复杂网络上作进一步尝试。

参考文献

[1] 王林,戴冠中.复杂网络中的社区发现 --理论与应用[J].科技导报,2005,23(8):62-66.

[2] NEWMAN M. Modularity and community structure in networks [J]. PNAS, 2006, 103 (23): 1-7.

[3] 杨格兰.基于复杂网络理论的产品结构模块划分方法[J].图学学报,2012,33(6):69-75.

[4] ZACHARY W W. An information flow model for conflict and fission in small groups [J]. Journal of Anthropological Research, 1977, 33: 452-473.

[5] JEONG H, TOMBOR B, ALBERT R, et al. The large-scale organization of metabolic networks [J]. Nature, 2000, 407: 651-654.

[6] POOL I, KOCHEN M. Contacts and influence [J]. Social Networks, 1978(1): 1-48.

[7] 贺德方,乔晓东,朱礼军,等.汉语科技词系统:新能源汽车卷[M].北京:科学技术文献出版社,2012.

[8] KERNIGHAN B W, LIN S. An efficient heuristic procedure for portioning graphs [J]. Bell System Technical Journal, 1970, 49: 291-307.

[9] 谢军.复杂网络中分析社团结构算法研究综述[J].信息通信,2010(4):48-51,71.

[10] GIRVAN M, NEWMAN M E J. Community structure in social and biological networks [J]. Proc of National Academy of Science, 2002, 9(12): 7821-7826.

[11] 刘大有,金弟,何东晓,等.复杂网络社区挖掘综述[J].计算机研究与发展,2013,50(10):2140-2154.

[12] TYLER J R, WILKINSON D M, HUBERMAN B A. Email as spectroscopy: Automated discovery of community structure within organizations [C]// Proc of the 1st Int Conf on Communities and Technologies. Amsterdam, Netherlands: Kluwer Academic Publishers, 2003: 81-96.

[13] PONS P, LATAPY M. Computing communities in large networks using random walks [J].Computer and Information Sciences, 2005, 3733: 284-293.

[14] 张娜.复杂网络社区结构划分算法研究[D].大连:大连理工大学,2009.

[15] NEWMAN M E J. Fast algorithm for detecting community structure in networks [J]. Physical Review E, 2004, 69(6): 66-133

[16] NEWMAN M E J, GIRVAN M. Finding and evaluating community structure in networks[J]. Physical Review E, 2004, 69(2): 26-113.

作者简介

殷希红，女，1988年生，硕士在读，研究方向：知识管理与技术，E-mail：yinxh2013@istic.ac.cn。
乔晓东，男，1965年生，研究员，研究方向：信息服务和信息资源。
张运良，男，1979年生，副研究员，研究方向：知识组织、自然语言处理。

Knowledge Organization System Concept Community Detection Based on Complex Networks

YIN XiHong, QIAO XiaoDong, ZHANG YunLiang
(Institute of Scientific and Technical Information of China, Beijing 100038, China)

Abstract: This paper applies the complex network theory into the representation of knowledge organization system, extracts the concept and concept relations from the knowledge organization system, and constructs undirected complex network and undirected weighted complex network. Using the walktrap community discovery algorithm, we find the concept community in complex network, in order to help users when they input seed concept, only to return the corresponding community. The modularity is used to evaluate the community discovery results, and to demonstrate the effectiveness of this method. This experiment regards the Chinese scientific and technical vocabulary system (new energy vehicles) as an example to discover the community, and finds that this method is fast and effective.

Keywords: Complex networks; Knowledge organization systems; Concept community; Community detection

（收稿日期：2014-07-03）