基于复杂网络聚类方法的网络社区发现技术研究

研究目的和意义

现实世界中的网络具有一些共同的拓扑统计性质，即“小世界性”和“无标度性”。这些性质既不同于规则网络，也不同于随机网络，正像近几十年来物理学家认为“复杂位于规则与随机之间”一样，所以大家把实际网络称为“复杂网络”。所谓小世界性是指实际网络具有比规则网络小得多的平均节点间距离和比随机网络大得多的平均集群系数（即邻点之间也相邻，形成紧密集团的比例）；而无标度性则指实际网络中节点邻边数取一个定值的概率分布函数是幂函数。这个幂函数标志基本单元与其邻居相互作用能力的极其不均匀分布。

复杂网络已经逐渐成为研究复杂系统的一种重要方法和跨学科的研究热点。随着对复杂网络性质的物理意义和数学特性的深入研究，人们发现许多时间的网络都具有一个共同性质——社区结构。社区结构是指网络是由若干个“群”或者“团”构成的。发现网络中的社区结构，对于了解网络结构与分析网络特性具有重要意义。社区结构分析在社会学、生物学、物理学、互联网和商业活动中都有广泛的应用。

社区发现是指在一个集合中，根据元素之间的某种关系将集合划分为若干社区（可交叉的子集）的过程。社区发现最基本的作用是将个体进行社区分类，将个体划分到若干社区中。划分得到的社区，可以为其成员提供个性化服务和信息推荐，也可用于以社区为活动单位的实际应用，如网络安全预警、社会行为分析等。

本课题利用聚类方法来研究复杂网络中的社区发现问题。课题组以国内高校教师的论文合作关系所形成的复杂网络为具体研究对象，以发现其中蕴含的研究社区，从而可以针对不同的社区用户提供个性化服务。这个网络将包含数以千计的节点，其节点规模要远远超过现在比较流行的网络社区发现的数据集，因此对社区发现算法的效率有更高的要求。因此，本课题的研究具有理论意义和实用价值。

国内外概况

网络是由许多节点与连接两个节点之间的一些边组成的，其中节点代表真实系统中不同的个体，而边则用来表示个体之间的关系。自然界中存在的大量复杂系统都可以通过形形色色的网络加以描述，如社会系统中的人际关系网、科学家协作网络和流行疾病传播网、基因调控网和蛋白质交互网，科技系统中的电话网、因特网和万维网等。由于这些网络具有很高的复杂性，因此被称为“复杂网络(complex network)”[1]。

最初对网络的研究来源于数学中的图论分支，随后社会学家也对网络开展了相关工作。社会学家比较典型的做法是发放调查问卷，询问调查对象与他人的交往关系，从而构建人与人之间的社会网络。在该网络中点代表个人，而边代表二者之间的交往关系。但是不管数学家还是社会学家之前的研究，都是集中于

只包含几十到几百个顶点的网络，在这样的网络中发现其特性。互联网的快速兴起及 Web 应用的发展不仅为网络的研究提供了更大的数据集，如包含众多科研人员发表论文信息的各种在线数据库，而且使 Web 应用本身形成了一定形式的网络，如万维网中各个 web 页面的相互链接组成的网络。特别是进入 Web 2.0 时代后，由大量网民参与的各种新的网络应用的兴起，使得人与人之间在网络上建立了更为直接的关系。博客之间的互引用关系、社会网络服务网站(SNS)的好友关系、购物网站中相同兴趣关系以及微博之间的转发关系等，这些关系组成了一个个庞大的网络，它们在一定程度上是用户真实社会关系的反映，包含了很多有用的信息。这些新的网络包含有大量的节点，节点之间又有着较为复杂的关系，不同于之前的各种研究。如果能通过一定的方法揭示其本质特性，挖掘出网络中有价值的信息，这必将极大地改变传统的信息获取方式，给用户带来更为精准、可信的数据，具有很强的现实意义和科学研究意义。如对用户购物关系的研究能够给用户推荐更为感兴趣的物品，对微博社会关系网络的研究能够更好的发现网络热点事件。目前对于这些社会网络的研究才刚刚开始，已经吸引了众多研究人员的关注，包括物理学家、生物学家和计算机科研人员都参与到其中。

1998 年，Watts 和 Strogatz 在《Nature》上发表论文阐述实际复杂网络的小世界效应(small-world effect)[2]，即节点间具有较小的平均路径长度，对数依赖于网络的规模。1999 年，Barabasi 和 Albert 在《Science》上发表论文指出，许多真实网络的度分布遵循幂律分布，称为无标度网络(scale-free network)[3,4]。而大量实证研究表明，真实网络不仅具有小世界和无标度等特性，还呈现明显的社区结构(community structure)[5,6,7]。

复杂网络可以建模为一个图 G=(V,E)，V 表示网络的节点集合，E 表示连接的集合。复杂网络可以使无向图，有向图，加权图或者超图来表示。所谓社区(community)，又称网络簇，是指网络中的稠密连通分支，具有簇内连接稠密、簇间连接相对稀疏的特点。社区结构是网络模块化与异质性的反映，表示真实网络可以看作是许多不同类型节点组合形成的。深入研究网络的社区结构不仅有助于揭示错综复杂的真实网络是怎样由许多相对独立而又互相关联的社区形成的，使人们更好的理解系统不同层次的结构和功能特性，而且具有重要的实用价值。社会网中的社区结构用于揭示具有共同兴趣、爱好或社会背景的社会团体；蛋白质网络中的社区结构可用于发现生物系统中功能相关的结构单元；万维网中的社区结构可用于提高网络搜索的性能和准确性，实现信息过滤、热点话题跟踪和网络情报分析等。因此，社区发现已成为复杂网络领域中的一个非常重要的研究方向。

Newman 将网络大致分为四类，包括社会网络(social networks)、信息网络(information networks)、技术网络(information networks)和生物网络(biological networks)[6]。本课题中，我们将科学家合作网络(coauthorship network)这一社会网络作为研究对象，进行社区发现的研究。

目前已经存在若干复杂网络聚类方法用于发现社区，按照所采用的基本求解策略，文献[1]将它们中的大多数归纳为两大类：基于优化的方法(optimization based method)和启发式方法(heuristic method)。前者将复杂网络聚类问题转化为优化问题，通过最优化预定义的目标函数来计算复杂网络的簇结构。后者将复杂网络聚类问题转化为预定义启发式规则的设计问题。文献[7]将社区发现算法分为图分割方法(graph

partitioning)和层次聚类方法(hierarchical clustering)。图分割算法主要被计算机科学家采用，而层次聚类方法多见于社会学方面，最近也被物理学家、生物学家及应用数学家应用到社会和生物网络中来。尽管这两类方法是为了解决同一个问题，但是二者之间却有明显的不同，图分割方法通常需要知道网络划分得到的社区的数量和大小，而层次聚类方法得到的社区数量由网络本身来决定。下面分别阐述每类方法的特点。

基于图分割的社区发现方法:

图分割方法的典型应用是计算机领域的并行计算,通常认为要为这类划分任务找一个确切的解决方案是一个 NP 完全问题，对于大规模的网络非常困难。经过多年努力，人们已经设计了很多启发式算法，并从中找到了一些好的解决方案，如谱方法。谱方法[7]采用二次型优化技术最小化预定义的“截”函数。当一个网络被划分为两个子网络时，“截”即指子网间的连接密度。具有最小截的划分被认为是最优的网络划分。针对不同的问题，提出了不同的截函数。谱方法本质上是一种二分法，在每次二分过程中，网络被分割成两个近似平衡的子网络，当网络中含有多个簇时，谱方法递归的分割现存的子网络，直到满足预先定义的停止条件为止。谱方法具有严密的数学理论，已发展成数据聚类的一种重要方法，广泛应用于图分割以及空间点聚类等领域。

Kernighan-Lin 算法(简称 KL 算法)[8]，通过基于贪婪优化的启发式过程把网络分解为两个规模已知的社区。该算法为网络的划分引入一个增益函数 Q，定义为落在两个社区内部的边数与落在两个社区之间的边数的差，然后寻找 Q 值最大的划分方法。KL 算法的主要缺点是必须明确地知道社区的大小，如果指定的值和实际的情况不一致，就会出现错误的社区结构划分。研究人员提出了许多启发式规则来优化 Q 值，包括贪婪方法、模拟退火方法、极值优化方法等，并取得了不错的结果。

Wu 和 Huberman 于 2003 年提出了一种基于电阻网络电压谱的快速谱分割方法(W-H 方法)[9]，基本思想是将两个不在同一社区内的节点看成源节点和终节点，将每条边视为一个阻值为 1 的电阻，那么在同一个社区内的节点之间的电压值应该是比较快接近的。因此只要通过正确的方法找到源节点和终节点，选择一个合适的电压阈值，就可以得到正确的社区结构。而且不需要事先确定网络中的所有社区结构，该方法即可以确定网络中的某一特定节点所属的社区。

基于层次聚类的社区发现方法

社会网络分析中的层次聚类方法的思想更接近社区结构的思想,目的是根据各种衡量节点之间的相似程度和节点之间的紧密程度的标准找出社会网络中的社区结构。层次聚类方法从具体实现上可以分为凝聚法和分裂法。凝聚法计算节点之间的相似度，并按照相似度由高到低的顺序，想本来为空的网络中添边，这个过程可以在任何时刻停止，并将在该时刻网络中的各组件作为社区。在凝聚法过程中，根据社区的不同划分方法，可进一步细分为单连接法和全连接法。分裂法的基本原理与凝聚法类似，只是其通过删除边进行的。一般做法是找出相互关联最弱的节点，并删除它们之间的边，通过这样的反复操作将网络划分为越来越小的组件，连通的网络构成社区。层次聚类方法被应用到各种各样的社会网络分析中。

由 Girvan 和 Newman 在 2002 年提出的 GN 算法也采用反复识别和删除粗间连接的策略聚类复杂网络[10]。社区之间的少数连接将成为社区间通信时通信流量的必经之路，如果考虑网络中某种形式的通信并

寻找到具有最高通信流量的边，则去除该边将获得网络最自然的分割。由此，Girvan 和 Newman 引入了边介数度量网络的通信流量，提出基于边介数的社区发现算法。基本思想是，迭代计算网络中每条边的介数并去除介数最大的边，直至网络中所有边被去除，每个点自成一个社区。GN 算法的最大缺点是计算速度慢，由于边介数的计算开销过大，GN 算法具有很高的时间复杂性，只适合处理中小规模的网络。尽管如此，GN 算法在复杂网络聚类研究中仍占有十分重要的地位，首次发现了复杂网络中普遍存在的网络簇结构，启发了其他研究者对这个问题的深入研究，掀起了复杂网络聚类的研究热潮。

Newman 等人引入了模块度(modularity)[11]来评价社区分解的合理性，其基本思想是，一个好的社区划分内部节点连接概率应远大于具有同样度序列的随机图中内部节点的连接概率。由于模块度的定义独立于特定的社区发现算法，社区发现问题可以简化为模块度优化问题。考虑到网络可能的划分方案数与网络规模成指数关系，对于大规模网络来说，穷尽搜索肯定不可行，人们引入了各种启发式的模块度优化方法，如贪婪算法[12,13]、极值优化方法[14]、模拟退火[15]等。

Son2006 年提出了铁磁随机域伊辛模型(FRFIM)进行社区发现，但是具有较高的时间复杂度，只适合于包含较少节点的网络[16]。Boccaletti 提出了另一个有效地社区发现方法，基于相位振荡器的聚类去同步化性质[17]。通过给每个网络节点分配一个合理的 N-维坐标，我们可以把网络聚类问题转换为传统的空间点聚类问题，然后采用 K-means 等经典聚类算法聚类这些新生成的空间点。该思想最早可以追溯到 1970 年 Hall 针对图分割问题提出的加权二次型变换变换算法[18],基于相似的思想，Donetti 和 Munoz 在 2004 年提出了一种结合谱方法和空间点聚类方法的复杂网络聚类算法[19]，首先通过计算拉普拉斯矩阵的 N 个最小特征向量将网络映射到 N-维空间中，然后采用某种基于距离的空间点聚类算法聚类网络节点。

其他社区发现方法：

针对基于连接的互联网聚类问题，Kleinberg 等人在 1999 年提出了著名的 HITS 算法[20]。该算法的基本假设是根据连接关系，互联网中存在权威和中心两种基本类型的页面，权威页面倾向于被多个中心页面引用，而中心页面倾向于引用多个权威页面。基于权威-中心页面间相互指向的连接关系，HITS 算法通过计算互联网对应的某些特殊矩阵的主特征向量来发现隐藏在互联网中的全部由权威-中心页面构成的网络簇结构。该算法被广泛地应用于包括 Altavista 在内的多个搜索引擎中。

2005 年 Palla 等在《Nature》上发表文章，提出了能够识别重叠网络簇结构的 CPM 算法[21]。该算法的基本思想是，网络簇由多个相邻的 k-团(k-clique)组成，相邻的两个 k-团至少共享 k-1 个节点，每个 k-团唯一的属于某个网络簇，但属于不同网络簇的 k-团可能共享某些节点。基于以上启发式信息，CPM 能够计算重叠网络簇结构，但是具有如下缺点：在实际应用中参数 K 难以确定，选取不同的 K 值往往得到差别较大的网络簇结构，但难以评判它们的优劣。

由于复杂网络的社区发现研究具有重要的理论意义和应用价值，不仅成为计算机领域中最具挑战性的基础性研究课题之一，也吸引了来自物理、数学、生物、社会学和复杂性科学等众多领域的研究者，掀起了一股研究热潮。从 2002 年至今，新的应用领域不断被拓展，《Nature》[15,21,22]、《Science》[23]、《Proceedings of National Academy of Sciences (PNAS)》[10,24,25]、《Physics Review Letter》[26,27]、《IEEE Transaction on

Knowledge and Data Engineering (TKDE) 》[28,29]、《PLOS Computational Biology》[30,31]等不同领域内的权威国际杂志和多个重要国际学术会议发表了很多相关论文。目前国内多家单位开展了复杂网络研究，如清华大学[32]，中国科技大学[33]、大连理工大学、中科院数学与系统研究所等，并在不同方面取得了很好的研究结果。

参考文献：

[1] 杨博,刘大有,Liu Jiming 等.复杂网络聚类方法.软件学报,2009,20(1):54-66.

[2] Watts DJ, Strogatz SH. Collective dynamics of small world networks. Nature, 1998, 393(4):440-442.

[3] Barabasi AL, Albert R. Emergence of scaling in random networks. Science, 1999, 286(5439):509-512.

[4] Barabasi A, Bonabeau E. Scale-Free networks. Scientific American, 2003, 288(5):60-69.

[5] Borgs C, Chayes J, Mahdian M, Saberi A. Exploring the community structure of newsgroup. IN: Kim W, Kohavi R, Gehrke J, DuMouchel W, eds. Proc. of the 10th ACM SIGKDD Int'1 Conf. on Knowledge Discovery and Data Mining. Seattle: AAAI, 2004. 783-787.

[6] Newman MEJ. The structure of scientific collaboration networks. Proc. of the National Academy of Science, 2001, 98(2):404-409.

[7] Newman MEJ. Modularity and community structure in networks. Proc. of the National Academy of Science, 2006, 103(23):8577-8582.

[8] Kernighan B W, S Lin. An efficient heuristic procedure for partitioning graphs. Bell. System Technical Journal, 1970, 49:291-307.

[9] Wu F, Huberman BA. Finding communities in linear time. A physics approach. European Physical Journal B, 2004, 38(2):331-338.

[10] Girvan M, Newman MEJ. Community structure in social and biological networks. Proc. of the National Academy of Science, 2002,9(12):7821-7826.

[11] Newman MEJ, Girvan M. Finding and evaluating community structure in networks. Physical Review E, 2004, 69(2). 026113.

[12] Newman MEJ. Fast algorithm for detecting community structure in networks. Physical Review E, 2004, 69(6). 066133.

[13] Clauset A, Newman MEJ, Moore C. Finding community structure in very large networks. Physical Review E, 2004, 70(6). 066111.

[14] Duch J, Arenas A. Community identification using extremal optimization. Physical Review E, 2005, 72(6). 027104.

[15] Guimera R, Amaral LAN. Functional cartography of complex metabolic networks. Nature, 2005, 433(7028): 895− 900.

[16] S. W. Son, el al. Random field Ising model and community structure in complex networks. The Euro. Physical Journal, 2006, 50(3): 431-437.

[17] Boccaletti S, Ivanchenko M, Latora V.et.al. Detection of complex networks modularity by dynamical clustering. Physical Review E, 2007, 75(4):045102.

[18] Hall KM. An r-dimensional quadratic placement algorithm. Management Science, 1970,17(3):219-229.

[19] Donetti L, Munoz Ma. Detecting network communities: A new systematic and efficient algorithm Journal of Statistical Mechanics:Theory and Expeiment, 2004, 10:P10012.

[20] Kleinberg JM. Authoritative sources in a hyperlinked environment. Journal of the ACM, 1999, 46(5):604-632.

[21] Palla G, Derenyi I, Farkas I, Vicsek T. Uncovering the overlapping community structures of complex networks in nature and society. Nature, 2005, 435(7043):814-818.

[22] Palla G, Barabasi AL, Vicsek T. Quantifying social group evolution. Nature, 2007, 446(7136):664-667.

[23] Ravasz E, Somera Al, Mongru DA. Hierarchical organization of modularity in metabolic networks. Science, 2002, 297(5586):1551-1555.

[24] Wilkinson DM, Huberman Ba. A method for finding communities of related genes. Proc. of the National Academy of Science, 2004,101(Suppl.1):5421-5428.

[25]Radicchi F, Castellano C, Cecconi F, Loreto V, Parisi D. Defining and identifying communities in networks. Proc. of the National Academy of Science, 2004, 101(9):2658-2663.

[26] Reichardt J, Bornholdt S.Detecting fuzzy community structures in complex networks with a potts model. Physical Review Letters, 2004, 93(19):218701.

[27] Garlaschelli D, Loffredo MI. Patterns of link reciprocity in directed networks. Physical Review Letters, 2004,93(26):268701.

[28] Yang B, Cheung WK, Liu J. Community mining from signed social networks. IEEE Trans. on Knowledge and Data Engineering,2007, 19(10):1333－ 1348.

[29] Brandes U, Delling D, Gaertler M, Görke R, Hoefer M, Nikoloski Z, Wagner D. On modularity clustering. IEEE Trans. on Knowledge and Data Engineering, 2008, 20(2):172－ 188.

[30] Cartozo CC, Rios PDL, Piazza F, Lio P. Bottleneck genes and community structure in the cell cycle network of S.pombe. PLOS Computational Biology, 2007, 3(6):e103.

[31]Wang Z, Zhang J. In serach of the biological significance of modular structures in protein networks. PLOS Computational Biology, 2007, 3(6): e107.

[32] 淦文燕，赫南，李德毅，王建民．一种基于拓扑势的网络社区发现方法．软件学报，2009，20(8): 2241-2254

[33] 赵鹏. 复杂网络与互联网个性化信息服务的研究.中国科技大学博士论文, 2006

研究内容，工作方案（包括研究方法、技术路线、进度安排、拟达到的技术指标、提交成果方式）。

研究内容

本课题利用聚类方法来研究复杂网络中的社区发现问题。课题组以国内高校教师的论文合作关系所形成的复杂网络为具体研究对象，以发现其中蕴含的研究社区的目标，从而可以针对不同的社区用户提供个性化服务，如兴趣小组、研究信息推送等。这个网络将包含数以千计的节点，其节点规模要远远超过现在比较流行的网络社区发现的数据集，因此对社区发现算法的效率有更高的要求。

为了实现上述研究目标，课题主要包括两方面的研究内容：

1）基于自然语言处理相关技术和信息抽取技术的高校教师论文合作关系网络自动构建。构建过程包

括教师个人信息的抽取和自动更新、论文信息的获取和更新、领域专家的跟踪等。该社会网络构建系统可以自动采集教师信息对其进行一个比较全面的展示，包括经过整合后的个人信息和研究信息。这可以为那些希望了解某教师信息的用户提供全方位、多角度、垂直化的服务。

2）面向大规模节点的网络社区发现方法研究。由于目前的网络社区算法的复杂度比较高，因此需要针对具体应用，融入领域知识，寻找能够对大规模节点网络进行网络社区发现的方法。

研究方法与技术路线

为了完成上述研究内容，本课题制定了如下的研究方法与技术路线，主要包括 3 各方面的内容，分别是基于信息抽取技术和自然语言处理技术的信息获取、基于高校教师论文合作关系的研究网络构建与基于改进 GN 方法的社区发现，下面分别对每一部分进行详细阐述。

1. 基于信息抽取技术和自然语言处理技术的信息获取

1) 教师个人信息收集及实验

课题组目前已收集全国重点院校 120 所共计 4134 名计算机相关专业的教师的信息，且数据仍在不断更新。每个教师包含的已有信息项如表 1 所示。

表 1 系统教师信息项

序号	信息项	说明
1	姓名	
2	性别	
3	职称	教授、讲师等
4	是否院士	是否为中科院或工程院院士
5	导师类型	硕士生导师、博士生导师
6	研究方向	教师的研究领域
7	电话	
8	电子邮件	
9	大学	教师所在的大学
10	个人页面链接	信息来源页面

在系统收集到的数据基础上，进行了教师个人信息的抽取实验。我们提出了一种基于块划分的网页信息抽取模型从中文教师页面中抽取感兴趣的教师信息项。在该模型中，首先对包含教师信息的 HTML 页面进行预处理，包括规范化 HTML 标签等，并进行分词和词性标注。然后利用具有语义划分作用的标点符号和 HTML 标签共同将整个页面分割为一个个句子，我们认为句子之间相互独立，即上一句的内容对当前句子内的信息抽取没有作用。条件随机域模型(Contional Random Fields, CRF)被作为标注模型对句子中的信息进行标注。

表 2.教师个人信息抽取实验中要抽取的信息项

信息类型	抽取任务	标记
基本信息	性别	G
	出生日期	BI
	导师类型	T
	职称	P

教育信息	本科毕业时间	BD
	本科学校	BU
	本科专业	BM
	硕士毕业时间	MD
	硕士学校	MU
	硕士专业	MM
	博士毕业时间	PD
	博士专业	PU
	博士学校	PM
联系方式信息	传真	F
	电话	PH
	邮箱	E

在实验中共使用了三类特征信息，词性特征，启发词特征和信息项值特征，详细介绍如下：

- 词性特征。使用词性标注得到结果。
- 启发词特征。对信息抽取具有很强启示作用的词语。例如，在教师职称信息的前面一般会有“职称”，电话信息的前面会有“电话”这样的词。通过对这类情况的分析，我们得到了 40 个有效词汇。
- 值特征。原始的中文单词本身被用来作为特征。此外，一些信息项，如“性别”，其值只能是“男”“女”之一，即来自于固定的一个值集合中。我们根据常识信息对这些信息项值进行收集，最终得到了 34 个值信息词汇。

为了对该方法进行评价，我们从已有系统中收集了 636 个教师网页，并进行人工标注作为实验数据集。每个教师网页所标记的信息项如表 2 所示。首先针对每一个词生成特征，发现对于基本信息和联系方式信息该方法已经能得到较好的结果。而教育背景信息，因为前后文之间的联系比较复杂，存在远距离控制等，信息抽取的结果不够好。为此，在接下来的实验中将有关教育背景的词特征扩展到句子级，实验表明有效提高了信息抽取的结果，如图 1 所示。

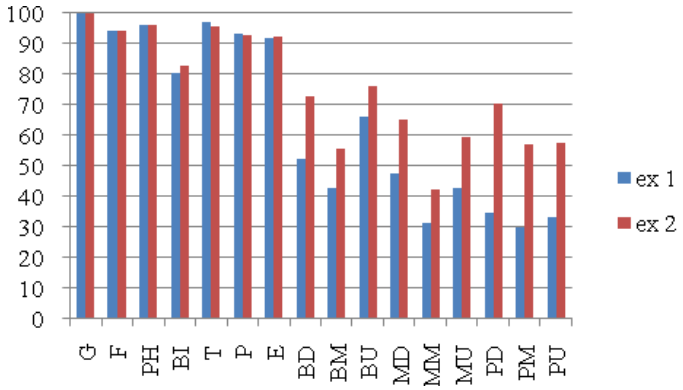

图 1. 基于 CRF 的教师基本信息抽取实验结果

2)论文信息收集

针对系统中已有的 4000 多个高校教师，系统从互联网上利用网络爬虫自动获取论文信息。具体流程如图 2。对于系统中已经收集到的每一个高校教师，把姓名和学校组合后作为关键词放到谷歌学术搜索引擎中进行检索，对于返回的检索结果，再到国内著名的学术网站万方[1]、知网[2]和维普资讯[3]进行查询，得到

[1] http://www.wanfangdata.com.cn/

更详细的论文信息并存储到数据库中。

2. 基于高校教师论文合作关系的研究网络构建

由于论文数据是从互联网上收集的，可能会存在错误信息。例如因为重名等原因，得到的论文信息可能是另外一个与其重名的教师的论文。因此需要从论文信息中去掉该部分杂质，提高信息的质量，以便下一步更好地构建社会网络。在之前获取论文详细信息的过程中，一部分论文能够得到作者的单位信息。我们假设每一个高校计算机学院的教师没有重名的，因此，可以通过作者单位信息对该论文进行一个初步的验证，判断是否该论文是当前查询教师的论文。再以得到验证的论文信息作为准确信息，对未得到验证的论文信息进行判断，从而得到较为准确的结果。

在构建社会网络时，将作者作为网络中的节点，而作者之间的合著关系则构成作者之间的连线。如果两个作者之间合作了一篇论文，那么很明显二者之间在现实社会中必然有着比较紧密的联系。通过从论文的合作者关系中抽取教师之间的联系，组建出与现实吻合的社会网络，然后找到该社会网络中联系比较紧密的小群体，即社区。整个过程如图 3 所示。

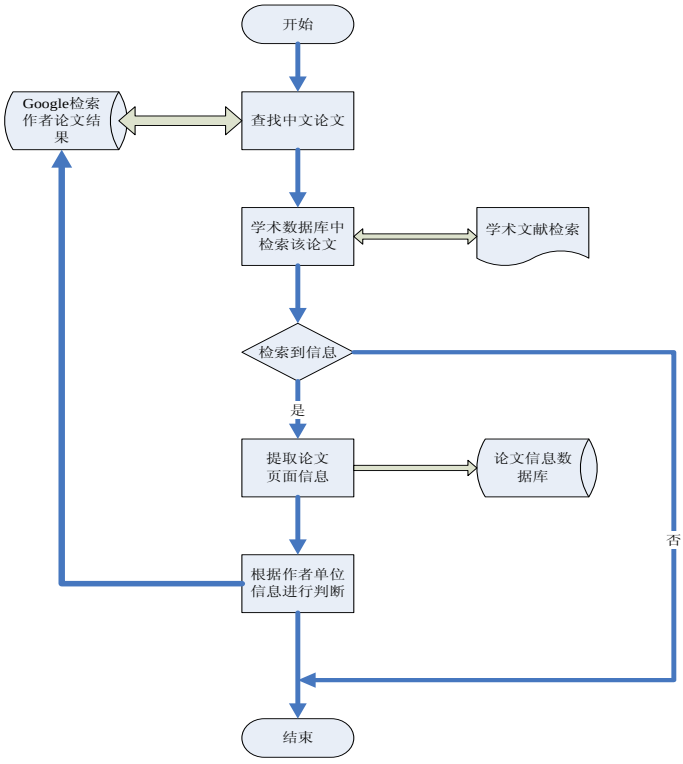

图 2. 论文获取流程

[2] http://www.cnki.net/
[3] http://www.cqvip.com/

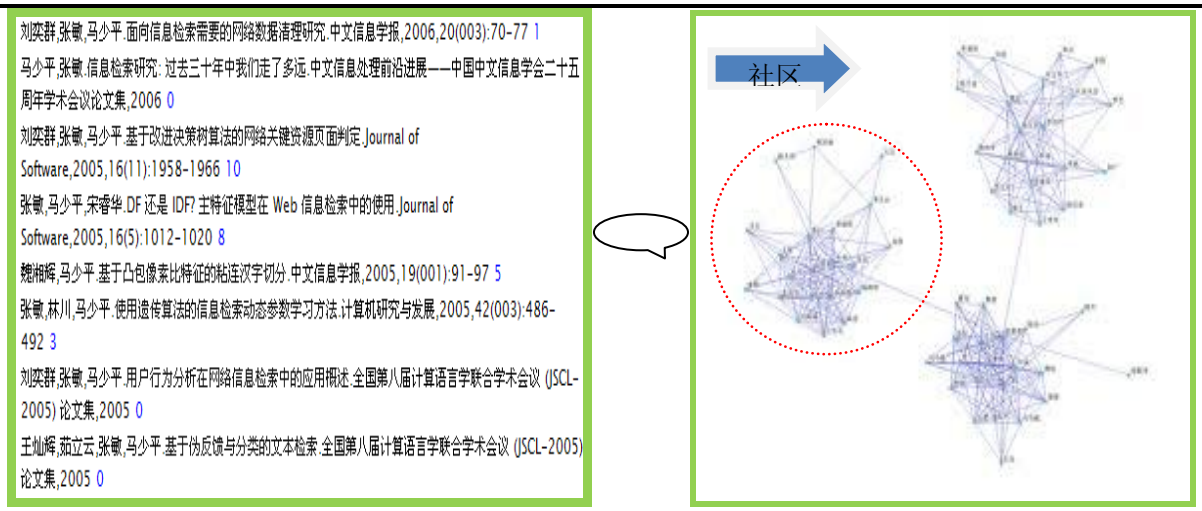

图 3. 社会网络构建过程

3. 基于改进 GN 方法的社区发现

Girvan 和 Newman 提出的 GN 算法是社区发现技术发展过程中的一个重要里程碑，它从网络的全局结构出发，避免了传统算法的若干缺点，成为目前进行网络社区分析的标准算法，得到了广发的应用。GN 算法的基本步骤如下：

step 1：计算网络中每条边的边介数
step 2：删除边介数值最高的边
step 3：重新计算所有边的边介数
step 4：重复第二步和第三步，直到所有的边都被删除

GN 算法包括计算网络中每条边的介数、去除边介数最大的那条边、重复进行直至网络中没有任何边存在。将该算法应用于某些网络，所有的结果与谱平分法得到的结果完全相同。对于一般的网络分析，GN 算法则具有较多的优越性。但是该算法也存在明显的缺点：首先，该算法无法预知网络最终应该分裂成多少社区；其次，计算速度缓慢，最差运行时间为 $O(m^2n)$，其中 m、n 分别为网路中的边数和节点数。为了克服以上缺点，研究人员提出了若干 GN 算法的改进方法。

Brandes 给出了一种快速重新计算边介数的方法，基本思想是：选择一个节点作为中心节点，只考虑中心节点和其他节点之间的最短路径，计算每条边由当前这些最短路径得到的边介数值，并将计算结果添加到当前该边的边介数之和中。然后，选择另外一个节点作为中心节点，并重复刚才的计算过程，直到每个节点都曾被选中作为中心点为止。在刚才的计算过程中，每条最短路径的端点都被计算过两次，计算得到的每条边的边介数的和正好等于该边准确的边介数的两倍。当节点集取的比较少时，它可以显著地提高计算速度，但同时降低了计算的准确性。Tyler 和 Wikinson 等在寻找电子邮件的社区结构和相关基因构成的社区时采用了该算法，取得了很好的社区划分结果。

本课题针对高校教师的论文合作关系进行社会网络分析，发现其中的社区。课题组计划选用 GN 算法

作为社区发现的基本方法。由于在我们的数据集中包含数千个节点，这对于 GN 算法来说是一件非常耗时的任务。特别是在我们运行的系统中，很多时候需要动态加入一些节点进来，如果每增加一个节点就重新运行 GN 算法进行一次社区发现，这是非常不可行的。为此，我们准备从以下三方面对 GN 算法进行改进：

1) 改善基本 GN 算法的时间复杂度。由上面的分析可知，基本 GN 算法的时间复杂度较高。而我们需要处理的高校教师社会网络包含有 4000 多个节点，如果直接用 GN 算法进行社区发现，需要耗费较长时间。为此，我们尝试引入领域知识(domain knowledge)，包括中文科研社会网络的常识特性以及作者的单位信息等知识，综合利用这些信息进行社区发现，减少算法的计算量，降低复杂度，同时改善结果的精确度。
2) 动态自适应社区发现。对于新加入网络中的节点，改进 GN 算法，不去重新计算所有节点的边介数，而是利用已有的社区情况，对新的节点进行社区分配，以快速的得到新的社会网络的社区划分结果，适应网络的动态扩张。
3) 社区成员重要性衡量。对于社会网络中发现出来的社区，如果能够对该社区内的不同成员的重要程度进行区分，给出衡量指标，是十分具有价值的。在社会网络中，每个人所处的地位和对网络的贡献度不尽相同，对于社区内处于中心地位的成员，其受关注程度也更高，人们对这些教师的查询行为也更多。如果能够识别出这些中心人物，就可以在后续工作中对他们的更多信息进行加工处理，更好的为用户提供这些重要人员的相关信息，具有很强的实用意义。