·复杂性科学·

基于LDA的复杂网络整体研究态势主题分析

赵紫娟[1]，李小珂[1]，郭 强[1]，杨 凯[1]，刘建国[2*]

(1. 上海理工大学复杂系统科学研究中心 上海 杨浦区 200093；2. 上海财经大学会计学院 上海 杨浦区 200433)

【摘要】复杂网络的研究发展非常迅速，已经对自动控制、统计物理、计算机及管理等学科产生了深刻的影响。然而，国内的主题发展态势一直缺乏系统、直观的分析。本文以2017年第十三届全国复杂网络大会的会议摘要文本为研究对象，从会议摘要主题分析的角度研究了国内复杂网络科研领域的整体发展态势。研究过程中首先对摘要文本进行预处理，通过建立自定义词典和停用词库对文本进行jieba分词，得到一个文档−词矩阵。然后用LDA主题模型对摘要主题进行挖掘，通过SVD分解确定主题数目，并基于摘要间的JS距离进行凝聚层次聚类，基于机构间的JS距离用Blondel算法对机构进行社团划分，最终得到10类会议主题和4类科研社团。实证结果不仅能分析出复杂网络宏观上的研究趋势与不同研究方向的热门程度；也能基于聚出的4类科研社团，为新进入复杂网络的研究者寻找对应研究方向的文献提供参考机构。

关 键 词 复杂网络； 社团结构； 研究态势； 文本分析

中图分类号 TP393； N949 **文献标志码** A doi:10.3969/j.issn.1001-0548.2019.06.019

Evolution Properties of Complex Networks in Terms of the LDA

ZHAO Zi-juan[1], LI Xiao-ke[1], GUO Qiang[1], YANG Kai[1], and LIU Jian-guo[2*]

(1. Research Center of Complex Systems Science, University of Shanghai for Science and Technology Yangpu Shanghai 200093;
2. School of Accounting, Shanghai University of Finance and Economics Yangpu Shanghai 200433)

Abstract The research of complex networks has been developing rapidly, which has had a profound impact on such disciplines as automatic control, statistical physics, computers, and management. However, there has been a lack of systematic and intuitive analysis of the development of topics in China. Taking the abstracts of the 13th National Complex Network Conference in 2017 as research object, we investigate the topic trend of the domestic complex network researches. Firstly, the text information of the abstracts are preprocessed and segmented by adding a custom dictionary and a stop word dictionary to obtain a document-word matrix. Then the LDA model is used to mine topics of the abstracts and SVD decomposition is applied to obtain the number of topics. As a result, ten topics of the conference are found through agglomerative hierarchical clustering according to the JS distance among the abstracts and four research communities involved in the conference are identified through community detection according to the JS distance among institutions. This work not only makes insight on the research trends and the popularity of different research directions in complex networks, but also provides reference institutions for new researchers to find corresponding research directions based on the results.

Key words complex networks; community structure; evolution properties; text mining

复杂网络是一门交叉性学科，近年来得到了大量来自不同领域学者的关注，在各个分支领域都有了丰硕的研究成果[1-5]。从宏观层面上，分析复杂网络的研究热点和研究趋势对于不同学科发展有着重要的意义。每年举行一届最具权威性的复杂网络大会，设置了复杂系统与复杂网络各个方面的讨论主题，吸引了来自国内外研究复杂网络学者的热情参与与投稿，包含了各个领域和方向的研究成果。如2017年全国复杂网络大会就有来自全国61个科研机构，投稿153篇论文摘要。研究分析这些最具前沿的科研成果，可以从一定程度上反映复杂网络目前的研究热点与方向。本文借助于收集到的复杂网络大会摘要的数据，利用文本分析的工具进行复杂网络研究态势的分析。

收稿日期：2018－06－19；修回日期：2018－12－03
基金项目：国家自然科学基金(61773248, 71771152)
作者简介：赵紫娟(1995－)，女，主要从事文本分析方面的研究.
通信作者：刘建国，E-mail: liujg004@ustc.edu.cn

目前研究者提出了较多的文本分析的方法，最早的经典主题模型方法是文献[6]在1990年提出的潜在语义分析(latent semantic analysis, LSA)方法。该方法使用词-文档矩阵，然后对该矩阵通过奇异值分解进行降维得到文本的主题，虽然解决了一词多义的问题，但是计算非常耗时，并且LSA得到的不是一个概率模型，缺乏统计基础，结果难以直观地解释。文献[7]提出了概率潜在语义分析(probabilistic latent semantic analysis, PLSA)，该方法基于统计学的理论，来分别估计文档-主题分布和主题-词分布，不过PLSA存在过拟合问题，对于新数据的适应能力不够。文献[8]在2003年提出隐含狄利克雷分布(latent Dirichlet allocation, LDA)，将PLSA贝叶斯化，即相比于PLSA固定的主题分布和词分布，LDA使用Dirichlet分布作为主题和词的先验分布，然后用吉布斯抽样求解后验分布，从而得到给定文档的主题分布。由于LDA有很好的适应性，在实际应用中LDA被应用到个性化推荐、广告预测等方面，是一种应用广泛的主题模型，因此本文选用LDA提取主题。

在主题模型中，首先需要确定主题数，大量实证研究证实LDA主题提取效果与文档主题数目K值有很大的关系，主题提取的结果对K值十分敏感[9]。文献[8]提出用困惑度(perplexity)定主题数目，但是这种方法会使主题数过大，产生主题冗余。文献[9]提出引入主题方差来决定主题数，用主题方差困惑度作为评定指标，其中困惑度为分子，主题方差为分母，然而这种方法对于主题之间差距不大的文本并不适用。文献[10]提出层次狄利克雷法(hierarchical Dirichlet processes, HDP)，是一种非参数贝叶斯模型，可以自主学习最优主题数目，其参数数目随样本数的增加而自适应，因而不需要提前决定主题数，不过HDP算法复杂度高，在文本分析中效率并不高。奇异值分解(singular value decomposition, SVD)可以将文档从高维空间映射到低维的潜在语义空间[11]，用保留的奇异值个数作为主题数，使得保留的矩阵能量信息不低于80%。本文在确定主题数K时，用SVD与困惑度方法做了对比，用困惑度确定K时一般需要从10～200取值，选取困惑度最小的K，需要训练至少20个主题模型，效率很低，并且通常产生的主题数偏大，因而本文选择用奇异值分解的方法确定主题数，这样选取的主题数不会过于冗余并且效率高。

本文选用LDA主题模型对会议摘要进行主题挖掘，利用SVD分解确定主题数目，以JS作为距离度量指标，对摘要进行层次聚类，用Blondel算法对机构进行社团划分，最后为每一类贴上标签。本文通过文本分析得出复杂网络宏观上的研究内容与不同研究方向的热门程度，并且通过对61个机构社团划分，得到这些机构所对应的研究方向。

1 模型与方法

本文收集了2017年第十三届全国复杂网络大会的投稿摘要信息，对其进行文本分析与主题建模，进而分析目前复杂网络的研究态势。另一方面，本文基于摘要的机构信息，结合摘要主题间的相似性，来分析机构间的关联关系，为研究者提供查询文献的参考建议。首先对初始文本进行预处理，转化为统一格式，再建立自定义词典和停用词库，利用jieba对文本进行切词处理；然后用SVD对词频矩阵分解来确定主题数，通过LDA主题模型对复杂网络大会摘要的主题进行发现；最后用JS距离度量摘要间的距离，对于摘要进行凝聚层次聚类，得到主题树状图，对于机构用Blondel算法进行社团划分，得到机构聚类结果图。分析过程如图1所示。

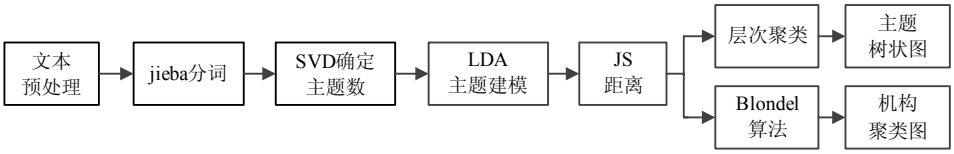

图1 数据分析流程图

1.1 LDA主题模型

LDA主题模型是PLSA算法的贝叶斯化模型，由文档、主题、词3层结构组成，是文本分析的有力工具，最常用于文本主题识别、文本分类以及文本相似度计算等方面。LDA采用的是词袋模型，将文档看作多个主题的混合分布，将主题看作不同词的混合分布，通过可观测到的文档-词分布估计文档-主题分布。LDA的图模型如图2所示。

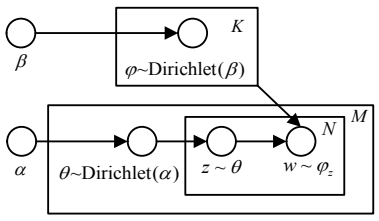

图2 LDA的图模型

其中，M表示训练的文档数目，每篇文档有N个词，K代表主题数，θ_i为文档d_i的主题分布，$\varphi_{z_{i,j}}$为主题$z_{i,j}$的词分布。Dirichlet(α)是参数为α的先验分布，供LDA采样生成文档对应的主题分布；Dirichlet(β)是参数为β的先验分布，模型中采样生成主题对应的词分布。LDA模拟了文档的生成过程，一篇文档先选定主题，根据主题选择词，不断重复选择主题和词从而得到文档，而LDA根据可以观测到的文档中的词不断训练得到了文档的主题以及主题分布。模型中所有变量间的关系表示如下：

$$p(w_i,z_i,\theta_i,\varphi|\alpha,\beta)=\prod_{i=1}^{N}p(w_{i,j}|\varphi_{z_{i,j}})p(z_{i,j}|\theta_i)\cdot p(\theta_i|\alpha)p(\varphi|\beta) \tag{1}$$

通过使用联合概率分布计算给定观测变量值下的隐含变量的条件分布[12]：

$$p(w_i|\alpha,\beta)=\int_{\theta_i}\int_{\varphi}\sum_{z_i}p(w_i,z_i,\theta_i,\varphi|\alpha,\beta) \tag{2}$$

在数据分析过程中，整个文档集作为输入内容进行LDA的训练，从而得到每篇摘要的主题分布。

1.2 SVD分解

LDA的主题数目没有一个固定的最优解，模型在训练时，主题数需要提前设定。本文选用奇异值分解来确定最佳主题个数。

奇异值分解(SVD)是一种矩阵分解方法，它将原始数据集矩阵$\boldsymbol{M}$分解成$\boldsymbol{U},\boldsymbol{\Sigma}$和$\boldsymbol{V}^{\mathrm{T}}$这3个矩阵。如果原始数据矩阵$\boldsymbol{M}$是$m$行$n$列，则$\boldsymbol{U},\boldsymbol{\Sigma}$和$\boldsymbol{V}^{\mathrm{T}}$这3个矩阵就分别是$m$行$m$列，$m$行$n$列和$n$行$n$列，$\boldsymbol{\Sigma}$是一个对角矩阵，矩阵$\boldsymbol{\Sigma}$中的对角元素$\delta$按从大到小顺序排列，这些对角元素称为奇异值，每个奇异值代表一个主题[13]。

K的取值公式为：

$$\sum_{i=1}^{j}\delta_i^2\Big/\sum_{i=1}^{m}\delta_i^2\geqslant\text{percentage} \tag{3}$$

式中，δ是奇异值；percentage是奇异值平方和占比的阈值；k远小于m和n。

1.3 聚类分析

1) 距离的度量

对于会议摘要语料进行LDA训练后，得到了153篇文档从Dirichlet分布中通过吉布斯采样出来的主题分布，因而选择JSD[14](Jensen-Shannon divergence)作为距离的度量指标，以下称为JS距离。JS距离是KLD(Kullback-Leibler divergence)的变形，它克服了KLD的无界和非对称的缺点，从信息熵的角度可以衡量相同时间空间里两个概率分布的差异情况。JS距离的取值范围为0～1，如果两个概率分布完全相同，则JS为0，当两个概率分布距离增大时，JS随之增大，最大为1。JS的计算公式如下：

$$D_{KL}(P\,\|\,Q)=\sum_i P(i)\ln\frac{P(i)}{Q(i)} \tag{4}$$

$$\mathrm{JSD}(P\|Q)=\frac{1}{2}D(P\|M)+\frac{1}{2}D(Q\|M) \tag{5}$$

式中，P和Q分别代表两篇文档主题的概率分布，$M=\frac{1}{2}(P+Q)$。

2) 层次聚类

层次聚类[15](hierarchical clustering)是无监督聚类算法，通过比较数据点间的距离，对所有数据点中最为相似的两个数据点首先进行合并，并反复迭代这一过程。本文通过计算JS距离度量文档之间的相似性，然后对文档进行凝聚层次聚类，并将聚类结果用树状图展示出来。

本文选择邓恩指数(Dunn validity index, DVI)[16]作为层次聚类的有效性评价指标：

$$\mathrm{DVI}=\frac{\min\limits_{0<m\neq n<K}\left\{\min\limits_{\substack{\forall x_i\in\Omega_m\\ \forall x_j\in\Omega_n}}\{\|x_i-x_j\|\}\right\}}{\max\limits_{0<m<K,\forall x_i,x_j\in\Omega_m}\max\{\|x_i-x_j\|\}} \tag{6}$$

式中，m和n为聚类后不同类别的样本容量；DVI是类间数据点的最短距离和类内数据点的最大距离的比值，最优的聚类结果需要有合适的类间聚类与合适的类内聚类，因而当DVI最大时，聚类效果最好。

3) Blondel社团划分算法

本文对机构所构建的网络进行社团结构划分，使用经典的Blondel方法[17]。Blondel基于文献[18]提出了一种用来衡量社团划分质量的指标——模块度来划分社团结构。该方法使用聚合思想，开始将每个节点看成一个社团，然后合并节点使得合并社团结构后的网络模块度最大化。不断迭代直到整个网络的模块度最大。

2 实证分析

2.1 文本预处理

文本预处理首先需要筛选过滤数据集，然后将数据集分别整理为基于摘要的主题分析数据和基于机构的聚类分析数据。为了提高切词效果，导入已经建立好的自定义词典和停用词库，利用Python的jieba分词包采用精确模式对文本数据进行切词分词

处理，进而得到文档-词矩阵。数据集的预处理流程图如图3所示。

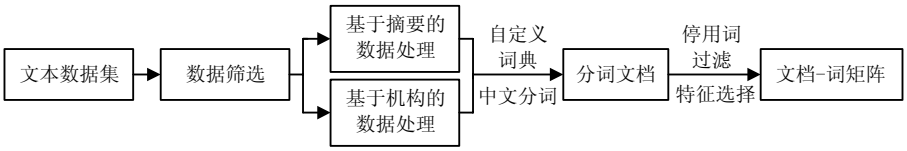

图3 预处理流程图

1) 数据集

本文收集了2017年11月24日～2017年11月27日在深圳举办的第十三届全国复杂网络大会中的155篇摘要文本。数据包含51篇中文摘要、103篇英语摘要和1篇内容未定摘要。摘要数据集所包含信息主要有：论文的标题、作者、作者所属的机构以及摘要内容。首先，本文对数据集进行预处理，删除了一篇不包含内容的摘要，对剩余154篇摘要进行去重，得到了102篇英文摘要和51篇中文摘要。考虑到摘要的作者99%均为中国人，并且jieba对于英文切词不是很理想，所以将英文摘要翻译为中文，最后将其合并，得到153篇摘要作为本文分析的数据。

根据后续工作的研究主体，本文将数据预处理分为两个方面：

① 针对摘要的主题分析，本文将102篇英文摘要翻译后与51篇中文摘要的内容进行合并，得到了153条总文本数据，作为摘要主题分析的数据集。

② 针对机构的主题分析，需要将同一机构的所有摘要合并为一行，得到61条以机构为单位的文本数据，作为机构主题分析的数据集。

2) 中文分词

在文本分析工作之前，首先对文本数据进行切词分词处理。本文用到的切词分词方法来自于Python开源的jieba分词包，将非结构化的描述性文本转化成结构化的数据。

尽管jieba有新词识别能力，但对于比较专业的术语jieba难免会错分。而本文所使用的文本数据是复杂网络专题的文本，包含有大量专业术语，因此为了保证更高的分词准确性，本文首先建立了自定义词典。在使用自定义词典之前，如摘要标题“网络中振幅死亡行为的牵制控制”会被分为“网络/中/振幅/死亡/行为/的/牵制/控制”。这样的分词结果使“振幅”、“死亡”、“牵制”、“控制”这几个词就失去了专业术语所包含的含义。因此，本文将“振幅死亡”和“牵制控制”添加到自定义词典中，然后再进行切词，就会得到“网络/中/振幅死亡/行为/的/牵制控制”，更加符合专业背景知识。

针对所有的摘要内容，最初构建了含有1 375个词条的自定义词典，并且在预处理过程中，对比原始语料和分词结果，不断完善自定义词典来提高分词的效果，最终得到了包含1 678个词的自定义词典。

另一方面，从分词结果中可以发现一些形容词、副词、量词、代词、连词、介词等出现频率高却没有包含学科内容，如“大量的、准确地、多个、我们、并且”等，为了减少干扰信息、计算时间和保证最终分词结果的精准性，本文将这些词进行了过滤。同时，摘要内容中类似于“提出、考虑、构建”等与复杂网络的主题没有关系的动词也进行了过滤，使得结果更加准确。为此，分析中建立了停用词库。在未使用停用词库时，如“根据老化特征对科学出版物进行排名”这句话，会被切分成“根据/老化特征/对/科学出版物/进行/排名”，而在引入停用词库后，分词结果会变成“老化特征/科学出版物/排名”，过滤掉了“根据”、“进行”等跟主题无关的词，使得分词结果更为准确。根据原始的文本数据以及实际研究需求，本文自建了不符合复杂网络主题相关的3 416个词条的停用词库，然后利用停用词库，对之前的分词结果进行过滤，得到了更为精简准确的分词结果。

2.2 摘要主题挖掘与分析

对于摘要文档-关键词的词频矩阵与机构文档-关键词的词频矩阵，进行SVD分解。主题数是进行奇异矩阵分解后得到的对角矩阵中奇异值保留的个数。目前存在许多方法指标确定奇异值的数目，最为常用的是确定保留的奇异值个数后使其所保留的奇异值的平方和能够达到矩阵总能量信息的80%～90%。在数据分析中，考虑到数据集为摘要文本，描述语言比较精简，故保留矩阵总能量信息的95%对应的奇异值。摘要-关键词矩阵与机构-关键词矩阵保留不同能量信息所对应的奇异值个数如表1和表2所示。

表1 摘要-关键词矩阵对应的奇异值个数

矩阵总能量信息比/%	80	85	90	95
奇异值个数	79	92	108	126

表2　机构-关键词矩阵对应的奇异值个数

矩阵总能量信息比/%	80	85	90	95
奇异值个数	27	31	38	46

在表1中，矩阵总能量信息占比为95%时，奇异值个数为126，因此设定摘要主题个数K=126。在表2中，矩阵总能量信息占比为95%对应的奇异值个数为46，因此设定机构主题个数K=46。对于经验参数α和β，参照文献[19]设置α=50/k，β=0.01。在LDA主题建模中，最关键的是要得到对两组参数的估计，分别为各主题的词分布φ和各文档的主题分布θ，LDA利用吉布斯采样不断训练迭代直至收敛来估计参数，本文分析中迭代次数设置为1 500，最后得到的参数估计如下：各主题下的词分布$\boldsymbol{\varphi}$表示各个主题下生成的每个词的概率，是一个$K\times V$的矩阵，其中V表示文档集中所有词的个数，K表示主题个数，针对摘要文档的K=126，针对机构文档的K=46。各文档下的主题分布$\boldsymbol{\theta}$表示每个文档中生成各个主题的概率，是一个$M\times K$的矩阵，其中M表示文档总数，针对摘要文档M=153，针对机构文档M=61。

表3和表4分别给出了经LDA模型训练后126个摘要主题和46个机构主题中的前4个。主题词后的数值越大表明该主题下产生该词的概率越大。

表3　基于摘要的Top4个主题

Topic 1th	数值	Topic 2nd	数值	Topic 3rd	数值	Topic 4th	数值
网络结构	0.640 5	智能电网	0.406 9	大数据	0.324 2	排名聚合	0.227 9
位置	0.093 7	智能体系统	0.057 9	物联网	0.256 1	教师	0.186 4
距离	0.070 3	一致性	0.056 9	云计算	0.142 1	学生	0.178 3
基准	0.070 3	控制	0.028 9	高性能	0.128 1	评价	0.064 1
晶格网络	0.023 5	指标	0.028 5	网络	0.038 1	排序算法	0.028 5
图分割	0.021 5	拓扑	0.028 2	随机效应	0.034 1	观察	0.021 4
人口数据	0.011 9	神经回路	0.015 2	数学家	0.024 1	有序	0.021 4
社区属性	0.011 8	离散时间	0.014 9	机遇	0.024 0	传播速率	0.021 4
觅食行为	0.011 8	监督	0.014 3	预期收益	0.014 1	大学	0.014 3
逻辑	0.011 5	优越性	0.014 1	网络科学	0.013 2	教务网	0.013 2

表4　基于机构的Top4个主题

Topic 1th	数值	Topic 2nd	数值	Topic 3rd	数值	Topic 4th	数值
网络建模	0.462 7	链路预测	0.393 4	同步性	0.276 1	网络节点	0.148 8
网络结构	0.159 6	联系	0.170 1	耦合矩阵	0.176 1	社区检测	0.040 7
动态	0.159 6	网络结构	0.165 4	内部	0.171 2	局部	0.032 6
吸引力	0.039 7	复杂网络	0.090 7	集体	0.065 5	纳什均衡	0.032 6
记忆	0.039 6	性能	0.071 4	股票市场	0.032 4	顶点	0.024 8
策略	0.033 2	游戏	0.051 4	信任	0.030 9	度	0.024 8
流行阈值	0.033 1	社区检测	0.042 0	评级	0.025 4	局部搜索	0.024 5
相互竞争	0.026 5	空模型	0.028 0	演化机制	0.024 0	聚类现象	0.024 5
免疫策略	0.026 4	信息	0.023 4	市场	0.020 3	测度函数	0.024 3
负相关	0.026 0	聚类现象	0.018 2	外部	0.020 1	游戏	0.016 3

从表3我们发现，摘要的前4个主题中“网络结构”、“智能电网”、“大数据”、“排名聚合”在每个主题下出现的概率最大，那么该摘要的主题分别跟这4个词的相关性最大。类似地，表4中机构的前4个主题中“网络建模”、“链路预测”、“同步性”、“网络节点”在每个主题下出现的概率最大，那么该机构的主题分别跟这4个词的相关性最大。

2.3　摘要聚类分析

1) 层次聚类结果及有效性评价

基于LDA得到的153篇摘要的文档-主题分布，本文采用JS距离计算摘要之间的距离，得到各个摘要间的相似性。本文利用层次聚类算法对距离最近的摘要进行聚合，最终在距离为1.24时聚为一类。本文用经典的聚类算法k-means与层次聚类结果对比，计算两种聚类方法的类内相似性与类间相似性加类内相似性的比值的均值，发现层次聚类的值要比k-means的大，说明本文适用于层次聚类算法。

实证中为了找到聚类效果最佳时的距离阈值，从聚类的最后一步倒推，当类内距离最小并且类间距离最大时，聚类效果最好。最后10步的聚类结果

如图4所示，图中黑色的横线是根据邓恩指数确定的，DVI=0.54时效果最好，分析中省略了摘要距离小于0.54的凝聚聚类过程。

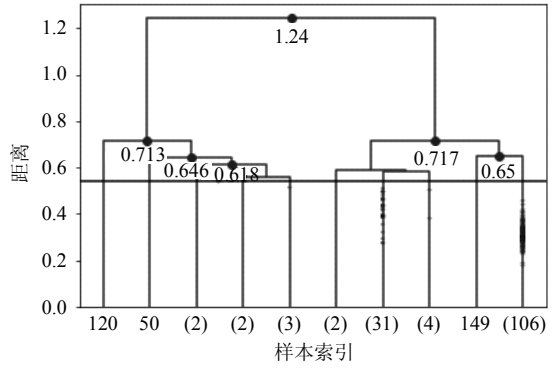

图4 最后10步聚类结果

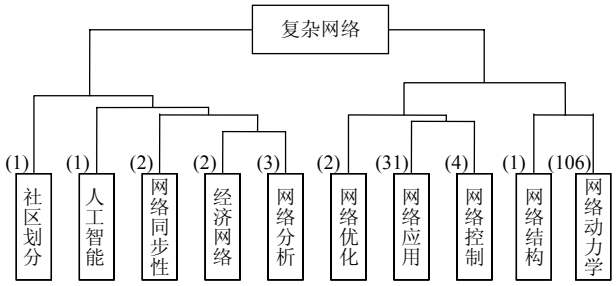

图5 主题树状图

2) 主题发现结果

根据图4的聚类结果提取10个类别所代表的主题，选出每一类别内摘要的主题词中词频最高的关键词归纳为该类的主题。这10个类别的样本容量分别为1、1、2、2、3、2、31、4、1和106。根据LDA的结果，网络动力学主题包含有网络同步性、博弈或传播等词；网络应用主题包含链路预测、推荐算法和命名游戏等；网络控制主题包含结构可控性、渗透阈值、控制器等词；网络分析主题包含网络安全、电路进行分析等词的摘要；单篇摘要为一类的有3篇，根据其具体内容细致归纳为人工智能、社区划分等。表5列举了排名前3的主题及其组成成分。从该表中可以看出，最热门的研究方向是网络动力学，其次是网络应用和网络控制。

根据图4的聚类结果和摘要的主题词，将10个类别的主题和最后10步聚类过程相结合，得到如图5所示的主题树状图。从图5中，不难发现原始数据集的153篇摘要被划分为了10类，分别为社区划分、人工智能、网络同步性、经济网络、网络分析、网络优化、网络应用、网络控制、网络结构和网络动力学。主题词上方的数字表示有多少篇摘要聚类到该主题下。从图中可以看到，这次复杂网络大会的研究方向主要有网络应用、网络控制、网络分析、网络应用等的，其中最热门的研究方向是网络动力学和网络应用。而不同研究方向彼此之间也有交集，如经济网络和网络分析在一定距离下聚为一类，网络结构和网络动力学在一定距离阈值下也聚为一类，所以科研人员可以将不同的研究方向相结合，拓展自己的研究方向，将不同方向的模型方法应用到自己的研究当中。

表5 摘要主题分布表

排序	摘要数	主题	组成成分
1	106	网络动力学	集群行为、社交网络、博弈、传播等
2	31	网络应用	链路预测、推荐算法、命名游戏等
3	4	网络控制	结构可控性、渗流阈值等

3) 机构聚类结果

本文对于153篇摘要按照其所属机构合并为61条数据进行数据分析，根据机构之间的JS距离，计算机构所投摘要的相似度，用Blondel算法对机构进行聚类，在调整阈值的过程中，为保证节点不被删除，阈值取值范围不超过每个节点与其他节点相似度的最大值，即保证每个节点至少有一条边存在。本文也用经典的聚类算法k-means与Blondel算法结果进行对比，计算两种聚类方法的类内相似性与类间相似性加类内相似性的比值的均值，Blondel算法的值要比k-means的大，说明这里Blondel算法进行聚类具有一定的优势。接着根据聚类结果建立邻接矩阵[20]，如果节点i和节点j属于同一类，则这两个节点之间有连边，否则两个节点之间没有边，生成基于摘要相似度的机构邻接矩阵，再用复杂网络分析软件Gephi进行可视化，机构聚类结果如图6所示。

对于每一个类中的机构，通过分析其摘要对应的主题词，总结出4个类别的研究方向如表6所示，在每个方向下有不同的研究内容。然而分析的数据仅为参与会议的机构所投摘要，不能全面地代表每个机构所有的研究方向。

表6 机构研究内容表

编号	机构数	主题	研究内容
1	14	网络应用	排名算法、链路预测、推荐算法、命名游戏等
2	10	网络控制	分布式控制、结构可控理论、渗流阈值等
3	23	网络动力学	社交网络、网络同步性、博弈、传播等
4	14	网络分析	链接分析、网络安全、电路分析等

如图6所示，61个机构聚为4类。第一类包含同济大学、国防科技大学、深圳大学等14个学校，这些学校的研究内容有链路预测、排名算法、命名游

戏等，基本包含在网络应用这个研究方向中；第二类有浙江大学、北京大学等10个学校，对应的研究内容有分布式控制系统、渗流阈值等，可以归纳为网络控制方向；第三类包含复旦大学、上海理工大学、华东理工大学、西安交通大学等23个学校，研究内容有集群行为、级联故障、社交网络等，研究方向可以归纳为网络动力学；第四类包含中山大学、上海交通大学、电子科技大学等14个学校，研究内容有链接分析、网络安全、电路分析等，可以归纳为网络分析这个研究方向。

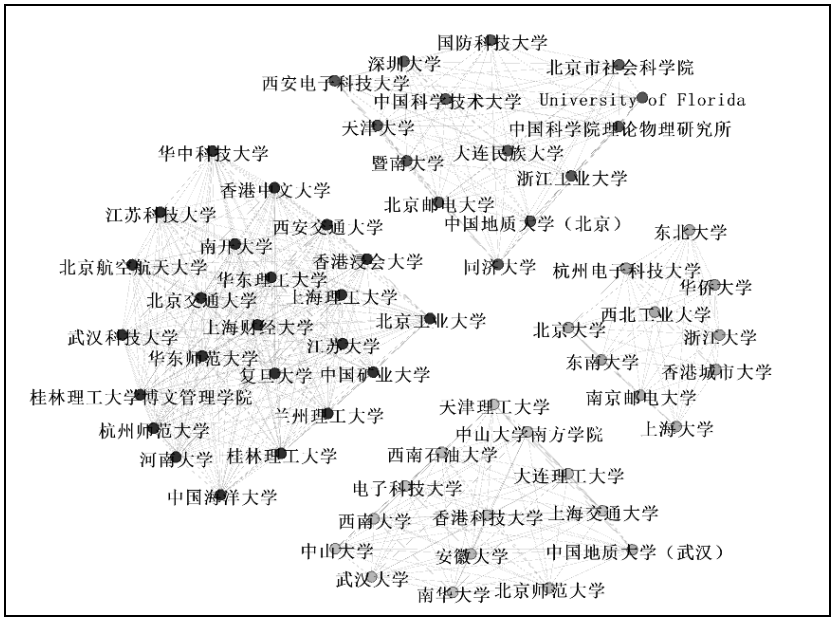

图6　机构聚类结果图

社团结构是复杂网络信息传播中最普遍和最重要的拓扑属性之一，本文通过分析机构的研究内容并进行社团划分，找到了研究方向相近的机构，从而为研究者提供合作的参考机构，对于研究知识传播具有重要意义。

3　结 束 语

本文基于2017年第十三届全国复杂网络大会的摘要数据，利用LDA模型提取摘要主题，通过SVD分解来确定主题个数，比困惑度方法更具有效率，且不会产生太大的冗余，得到了摘要的文档-主题矩阵，利用JS算法计算摘要间的距离，进一步基于摘要的JS距离进行凝聚层次聚类，得到主题树状图，分析复杂网络的研究态势。通过数据分析得出10类主题，分别为：网络动力学、网络结构、网络控制、网络应用、网络优化、网络分析、经济网络、网络同步性、人工智能和社区划分，其中网络动力学和网络应用为热门研究方向。另一方面，将机构作为研究主体，同样地，利用机构的文档-主题矩阵，使用JS算法计算机构间的距离，然后用Blondel算法对机构进行社团结构划分，得到机构的聚类结果。本文将参与会议的机构划分为4个社团，每个社团的研究方向分别为：网络应用、网络控制、网络动力学和网络分析。

本文通过对复杂网络会议文本进行研究，挖掘出复杂网络当前的研究趋势，可以帮助复杂网络的研究人员了解复杂网络学科最新的热门领域，拓展他们的科研方向，同时为复杂网络新的研究者提供宏观层面的认识，方便他们选择感兴趣的方向。还能基于机构聚类结果，为新的研究者提供依据机构寻找科研文献的参考建议。此外，本文也存在一些不足，如：自定义词典以及聚类后的主题归纳都受主观因素的影响，人工归纳标签的好坏还没有找到合适的评价指标；分析数据为参与会议的机构所投摘要，不能全面地代表各个机构所有的研究方向。本文还有进一步可扩展的工作：文本主题数的确定和聚类方法的选取都可以尝试更多的方法，也可以结合主题发现结果和机构聚类结果做科研合作单位的推荐。

参 考 文 献

[1] NEWMAN M E J. The structure and function of complex networks[J]. SIAM Review, 2003, 45(2): 167-256.

[2] LIU J G, LEI H, XUE P, et al. Stability of similarity measurements for bipartite networks[J]. Scientific Reports, 2016, 6: 18653.

[3] LIU J G, LIN J H, GUO Q, et al. Locating influential nodes

via dynamics-sensitive centrality[J]. Scientific Reports, 2016, 6(3): 032812.

[4] YANG K, GUO Q, LI S N, et al. Evolution properties of the community members for dynamic networks[J]. Physics Letters A, 2017, 381(11): 970-975.

[5] BARABÁSI A L, ALBERT R. Emergence of scaling in random networks[J]. Science, 1999, 286(5439): 509-512.

[6] DEERWESTER S, DUMAS S T, FURNAS G W, et al. Indexing by Latent semantic analysis[J]. Journal of the American Society for Information Science, 1990, 41(6): 391-407.

[7] HOFMANN T. Probabilistic latent semantic analysis[C]// The 15th Conference on Uncertainty in Artificial Intelligence. [S.l.]: Morgan Kaufmann Publishers Inc, 1999: 289-296.

[8] BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 601-608.

[9] 关鹏, 王曰芬. 科技情报分析中LDA主题模型最优主题数确定方法研究[J]. 现代图书情报技术, 2016, 32(9): 42-50.
GUAN Peng, WANG Yue-fen. Identifying optimal topic numbers from Sci-Tech information with LDA model[J]. New Technology of Library and Information Service, 2016, 32(9): 42-50.

[10] TEH Y, JORDAN M, BEAl M, et al. Hierarchical Dirichlet processes[J]. Journal of the American Statistical Association, 2007, 101(476): 1566-1581.

[11] 吴志祥, 王昊, 王雪颖, 等. 基于奇异值分解的专利术语层次关系解析研究[J]. 情报学报, 2017, 36(5): 473-483.
WU Zhi-xiang, WANG Hao, WANG Xue-ying, et al. Study on Chinese patent terms hierarchy parse based on singular value decomposition[J]. Journal of the China Society for Scientific and Technical Information, 2017, 36(5): 473-483.

[12] 曹娟, 张勇东, 李锦涛, 等. 一种基于密度的自适应最优LDA模型选择方法[J]. 计算机学报, 2008, 31(10): 1780-1787.
CAO Juan, ZHANG Yong-dong, LI Jin-tao, et al. A method of adaptively selecting best LDA model based on density[J]. Chinese Journal of Computers, 2008, 31(10): 1780-1787.

[13] 张俊博, 李健, 张宏宇. 潜在语义分析中主题数的确定方法[J]. 信息技术, 2016(7): 96-100.
ZHANG Jun-bo, LI Jian, ZHANG Hong-yu. Determination method of the number of topics in latent semantic analysis[J]. Information Technology, 2016(7): 96-100.

[14] MAJTEY A P, LAMBERTI P W, PRATO D P. Jensen-Shannon divergence as a measure of distinguishability between mixed quantum states[J]. Physical Review A, 2005, 72(5): 762-776.

[15] JOHNSON S C. Hierarchical clustering schemes[J]. Psychometrika, 1967, 32(3): 241-254.

[16] DUNN J C. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters[J]. Journal of Cybernetics, 1973, 3(3): 32-57.

[17] NEWMAN M E J, GIRVAN M. Finding and evaluating community structure in networks[J]. Physical Review E, 2004, 69(2): 026113.

[18] BLONDEL V D, GUILLAUME J L, LAMBIOTTE R, et al. Fast unfolding of communities in large networks[J]. Journal of Statistical Mechanics, 2008(10): 155-168.

[19] GRIFFITHS T L, STEYVERS M. Finding scientific topics[J]. Proc Natl Acad Sci USA, 2004, 101 (sup 1): 5228-5235.

[20] 汪小帆, 李翔, 陈关荣. 网络科学导论[M]. 北京: 高等教育出版社, 2012.
WANG Xiao-fan, LI Xiang, CHEN Guan-rong. Network science: An introduction[M]. Beijing: Higher Education Press, 2012.

编 辑 蒋 晓