doi:10.3969/j.issn.1000-436x.2014.01.011

基于扩散小波的社会网络多尺度分析

申国伟[1]，杨武[1]，黄文廷[2]，王巍[1]，丁丽[2]，于淼[1]
（1. 哈尔滨工程大学 信息安全研究中心，黑龙江 哈尔滨 150001；2. 国家计算机网络应急技术处理协调中心，北京 100029）

摘 要：针对社会网络中的同质性关系提出一个基于扩散小波的多尺度分析框架，通过局部相似性度量构造扩散算子，在统一的框架下对社会网络中的结构、内容、用户行为等进行多尺度分析。在合成和真实网络数据上进行实验，与典型算法的对比表明，本算法在无参数的条件下快速收敛并得到更好的结果。
关键词：社会网络；多尺度；同质性；扩散小波；扩散算子
中图分类号：TP393 **文献标识码**：A **文章编号**：1000-436X(2014)01-0089-10

Multi-scale analysis of social network based on diffusion wavelets

SHEN Guo-wei[1], YANG Wu[1], HUANG Wen-ting[2], WANG Wei[1], DING Li[2], YU Miao[1]
(1. Research Center of Information Security, Harbin Engineering University, Harbin 150001, China;
2. National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing 100029, China)

Abstract: A new multi-scale framework based on diffusion wavelets was proposed to analyze the homogeneous relationships, which can be used to conduct multi-scale analysis on structures, contents and user behaviors. The diffusion operator of diffusion wavelet only considers the local similarity in this framework. The experiments on both synthetic and real-world networks show that the proposed algorithm outperforms the typical algorithms in multi-scale analysis without parameters.
Key words: social network; multi-scale; homogeneity; diffusion wavelets; diffusion operator

1 引言

随着社交网络的兴起，人们的沟通和交流变得日益紧密。受共同兴趣的影响，人们在社会网络中产生显式或隐式的关系，如朋友关系、合作关系、消息传播关系、用户之间的交互行为关系等。其中，朋友关系和合作者关系具有明显的社区结构特征，用户发布的消息表现出话题相似性关系，用户在社会网络中的行为表现出趋同性，这些关系都是在同类实体之间产生，因此满足同质性特征[1]。由于社会网络演化速度快，在不同的尺度下具有不同的特征，很难在单一尺度下准确描述网络的关系特征。因此，本文针对社会网络中表现出来的同质性特征进行多尺度建模分析。

目前，社会网络分析的主要对象包括结构、内容、用户行为等，其中结构分析是社会网络研究最多的一个主题[2]，而且动态社区、多尺度社区发现成为当前结构分析的热点。经典的社区划分算法包括基于模块度的社区发现算法[3]、基于谱分析的社区发现算法等。Newman 等人在模块度方法的基础上提出模块度矩阵[4]，利用该矩阵的谱信息进行社区划分。模块度矩阵的提出将模块度方法与谱方法有效地结合起来，但是基于模块度的优化方法存在解析度限制[5]及时间复杂度高的问题。由于真实数据的多尺度结构特征，需引入多尺度社区划分算法。Mucha 等人在模块度社区发现算法的基础上提出了基于时态的多尺度社区结构发现算法[6]，同样通过分析链接可揭示多尺度特征[7]。Martelot 等利用全局和局部信息提出了快速的多尺度社区划分算法[8]，但是其不能处理有向网络。国内的程学旗

收稿日期：2013-05-02；**修回日期**：2013-11-18
基金项目：国家高技术研究发展计划（“863”计划）基金资助项目（2012AA012802）；国家自然科学基金资助项目（61170242）；中央高校基本科研业务费专项基金资助项目（HEUCF100614）
Foundation Items: The National High Technology Research and Development Program of China (863 Program) (2012AA012802); The National Natural Science Foundation of China (61170242); The Fundamental Research Funds for the Central Universities (HEUCF100614)

等利用扩散过程中的局部平衡态识别多尺度社区结构[9]。针对节点度异质的网络，在降维分析的基础上提出了利用协方差矩阵的谱隙来识别不同的拓扑尺度[10]，但是该方法受到协方差矩阵及主成份分析的限制。另外随着社会网络规模的增加，计算时间复杂度成为分析的瓶颈，而通过构建基于局部相似性[11]的扩散算子能够有效地处理大规模网络。

除了结构分析外，面向内容的话题分析也是数据挖掘、社会计算的核心任务之一。经典的话题模型为 LDA[12]，在该模型的基础上又出现了大量针对该模型的推广及改进。已有大量的文献针对微博网络中的内容进行分析，着重于微博与传统媒体对应的话题模型的区别，并提出了一系列改进话题模型[13,14]。Wang Chang 等人提出利用扩散模型处理文本数据集，实现多尺度主题分析[15]，但是在处理短文本消息时准确性明显下降。微博消息中的用户属性特征可用于提高多尺度话题分析的准确性。

用户行为分析对于理解用户行为模式、设计高效的沟通平台都起着重要作用[16]。Flavio Chierichetti 等人将用户在网络中的行为建立成马尔可夫模型进行分析[17]，但是社会网络中的用户行为变化非常迅速，传统的方法很难捕捉到用户的细粒度交互过程。社会网络中的用户行为可以看成随机游走过程，但是用户之间受到共同兴趣的影响表现出聚集现象。因此，可以结合随机游走理论和扩散小波分析用户交互行为在不同时间尺度下的动态变化过程。

综上所述，随着社会网络的规模增大、复杂性提高，必须充分考虑规模及动态性对社会网络分析的影响[18]。通过降维分析能够有效的处理复杂网络数据，主要包括主成份分析(PCA)[19]、特征映射(Isomap)[20]、拉普拉斯特征映射(laplacian eigenmaps)[21]等，而这些降维方法存在仍需对降维之后的低维空间进行分析或者无法直接处理有向网络的问题。因此，本文通过局部相似性度量社会网络中的关系强度，并在此基础上提出采用基于扩散小波[22]的框架实现多尺度降维分析[23]。由于扩散小波针对扩散算子进行操作，在降维分析的同时能够描述出数据的多尺度特征。

本文的主要工作有两点：一是针对社会网络中关系的同质性特征进行分析，提出了基于扩散小波的多尺度分析框架，并对该框架进行了理论分析；二是针对社会网络分析中社区结构发现、主题发现、用户行为分析 3 个主题给出了其局部相似性度量方法及实验验证。

2 社会网络多尺度分析框架

本节面向社会网络中同质性关系提出一个统一的多尺度分析框架，介绍框架中的描述模型、相似性度量方法及扩散矩阵的构造。

2.1 基于扩散小波的多尺度建模框架

本文针对大规模社会网络的同质性关系分析提出了一个基于扩散小波的多尺度分析框架 MSA-DW (multi-scale analysis based on diffusion wavelet)，该框架主要包括如下。

1) 针对数据集 X，将同质关系建模成有向带权图 $G=(V,E,W)$及关系矩阵 $\boldsymbol{A}$。V 为图中的顶点集合，E 代表图中的边的集合，W 为图 G 中边的权值。通过图 G 构建关系矩阵 $\boldsymbol{A}$，每一列相当于数据集中的一个元素的多维特征描述。

2) 计算图 G 中任意 2 个节点 i、j 之间的相似性 $S_{i,j}$，并构造相似性度量矩阵 $\boldsymbol{S}$。

3) 通过 $\boldsymbol{S}$ 计算节点 i、j 之间的局部扩散距离 $d(i,j)$，并构造扩散算子及对应的扩散矩阵 $\boldsymbol{T}$。

4) 构造扩散小波，分析扩散矩阵 $\boldsymbol{T}$。

5) 利用第 4）步中分析得到的扩展基描述出多尺度结构特征。

该框架中的所有待分析数据都采用有向带权图 G 进行描述，其中 V 可以是用户或者词等，W 即为任意两者之间的同质性关系强度。对于框架中的第 2）步，可以针对不同的分析任务采用对应的相似性度量标准，具有良好的可扩展性。

本文提出的分析框架 MSA-DW 主要具有以下几个优点。

1) 在同一个框架下可对社会网络中的结构、内容、用户行为等进行多尺度分析。

2) 扩散小波针对包含局部相似性结构特征的扩散矩阵进行操作，在降维分析的同时描述出多尺度的结构特征。

3) 无需对有向网络对应的描述矩阵进行对称化处理。

4) 无需人工设置尺度参数，尺度参数完全根据待分析数据来确定，在扩散小波计算过程中会自动收敛。

2.2 社会网络描述模型

对社会网络的关系特征建模是分析社会网络的基础，本节给出了社会网络中的结构、话题、动

态交互行为描述模型。

2.2.1 *结构关系描述模型*

社会网络中用户之间包含多类结构关系，例如人人网的朋友关系，新浪微博网络中的关注关系等。对于数据集 X 为用户的结构关系集合时，通过有向带权图 $G=(V,E,W)$描述，其中 V 代表网络中的用户，E 为用户之间的结构关系，W 为关系强度。如果用户 i 和 j 之间存在关系，其强度描述为 $W(i,j)$。通过图 G 构建关系矩阵 $\boldsymbol{A}$，其中 $\boldsymbol{A}(i,j)=W(i,j)$，关系矩阵 $\boldsymbol{A}$ 中的每一列相当于每一个用户的多维关系特征描述。

有向图中 $W(i,j)$不一定等于 $W(j,i)$，因此其对应的关系矩阵可能为非对称矩阵，已有的研究方法都是采取将其转换成无向图进行处理，但是这样会失去数据本身的结构特征。本文无需对非对称矩阵进行对称化操作，可有效地保护数据的结构信息。

2.2.2 *话题描述模型*

如果数据集 X 为用户发布的消息集合，对消息进行词切分，可以构建词—消息的关系矩阵 $\boldsymbol{M}$，$\boldsymbol{M}(i,j)$ 表示第 i 个词在第 j 条消息中的权重。对于词—消息矩阵 $\boldsymbol{M}$ 来说，该矩阵为一个非对称矩阵。矩阵 $\boldsymbol{M}$ 对应的奇异值按照降序排列为：$\sigma_1>\sigma_2>\cdots>\sigma_r$，即矩阵 $\boldsymbol{M}$ 的秩 Rank($\boldsymbol{M}$)=r，可将矩阵 $\boldsymbol{M}$ 分解成 $\boldsymbol{M}=\boldsymbol{U}\sum\boldsymbol{V}^{\mathrm{T}}$，其中 $\sum=\mathrm{diag}(\sigma_1,\sigma_2\cdots\sigma_r)$，$\boldsymbol{U}$ 为矩阵 $\boldsymbol{M}$ 的左特征矩阵，$\boldsymbol{V}$ 为 $\boldsymbol{M}$ 的右上三角矩阵，因此可以计算得到关系矩阵 $\boldsymbol{A}$

$$\begin{aligned}\boldsymbol{A}&=\boldsymbol{M}\boldsymbol{M}^{\mathrm{T}}=(\boldsymbol{U}\sum\boldsymbol{V}^{\mathrm{T}})(\boldsymbol{U}\sum V^{\mathrm{T}})^{\mathrm{T}}\\&=\boldsymbol{U}\sum\boldsymbol{V}^{\mathrm{T}}g\boldsymbol{V}\sum{}^{\mathrm{T}}\boldsymbol{U}^{\mathrm{T}}\\&=\boldsymbol{U}\sum\sum{}^{\mathrm{T}}\boldsymbol{U}^{\mathrm{T}}\end{aligned}\tag{1}$$

对于数据集 X，对消息进行词切分，构建词—词之间的关系图 G，其对应关系矩阵为 $\boldsymbol{A}$。则式(1)中矩阵 $\boldsymbol{U}$的每一个列向量即为关系矩阵 $\boldsymbol{A}$的特征向量，因此，可以基于关系矩阵 $\boldsymbol{A}$ 进行主题分析[15]。

社会网络中的文本大多数属于短文本，特别是微博网络中的消息不超过 140 个字，传统的话题模型在针对短文本分析时效果并不理想。本文针对微博短文本消息主题分析，在构建词—消息矩阵 $\boldsymbol{M}$ 时不仅考虑微博消息的短文本特性，而且考虑用户之间的互动关系。因此，计算 $\boldsymbol{M}(i,j)$时，除了考虑词频，同时需要考虑用户的历史属性特征等。

2.2.3 *动态交互行为描述模型*

通过结构分析可以发现，社区内的用户之间交互非常频繁，属于一种强关系，而社区间的用户产生的交互属于一种弱关系。用户之间通过交互表现出聚集特征，因此，交互过程分析成为社会网络动态分析的基础。

用户之间的交互行为在整体上表现出趋同现象，在不同的时间尺度下表现出不同的聚集特征。本文通过活动图 $G=(V,E,W,t)$ 来描述用户在社会网络中的动态交互过程。V 为用户集合，$E(i,j)$为 t 时刻用户 i 向用户 j 发起交互行为产生的有向边，W 为交互的权重。

在用户活动图 G 上构建有限状态的马尔可夫链，其中转移状态数为$|E|$，所有转移状态的集合 V 记为状态空间。对于 $\forall t>0$，状态空间 V 中的 $i,j,i_0,i_1,\cdots$，可以将单步转移过程描述为

$$\begin{aligned}p_{i,j}&=P\{V_{t+1}=j\,|\,V_0=i_0,V_1=i_1,\cdots V_t=i\}\\&=P\{V_{t+1}=j\,|\,V_t=i\}\end{aligned}\tag{2}$$

$P\{V_{t+1}=j\,|\,V_t=i\}$ 为从 i 到 j 的单步转移概率，因此，t 步转移概率则记为

$$p_{i,j}^{(t)}=P\{V_{i+t}=j\,|\,V_i=i\}\tag{3}$$

通过马尔科夫链构建关系矩阵 $\boldsymbol{A}$，其中 $A(i,j)=p_{i,j}$，关系矩阵 $\boldsymbol{A}$ 为一个非对称矩阵。

2.3 扩散算子的构造

分析框架中的扩散矩阵描述了数据的局部结构特征，所以扩散算子的定义在该框架中起着重要作用。本节在描述模型的基础上定义局部相似性扩散算子，并构建扩散矩阵。

2.3.1 *相似性度量*

图 G 中的任意的 2 个节点 i、j，其相似性度量为 $S_{i,j}$，满足 $S_{i,j}>0$。目前，相似性度量方法主要包括局部相似性度量和全局相似性度量方法。其中常见的相似性度量方法为基于高斯核的相似性度量，对于图 G 及其对应的关系矩阵 $\boldsymbol{A}$，节点 i、j 对应的特征向量分别为 $\boldsymbol{I}$、$\boldsymbol{J}$，即关系矩阵 $\boldsymbol{A}$ 中对应的列向量。因此，基于高斯核的相似性度量定义为

$$S_{i,j}=\exp\left(-\frac{\|I-J\|^2}{2\sigma^2}\right)\tag{4}$$

该核函数包含参数 σ，其对相似性度量的影响较大，但是该参数完全可以根据数据集的特征确定，不需要任何先验知识。

文献[9]对相似性度量方法进行了总结，但本文中只考虑局部相似性度量。对于最简单的相似性度量即为节点 i、j 共同拥有的邻域 C，这也是话题相

似性度量的一种有效方法。对于 $\forall i \in X, \forall j \in X$，其邻域分别为 $\Gamma(i)$、$\Gamma(j)$，相似度量为

$$S_{i,j} = \left|\Gamma(i) \cap \Gamma(j)\right| \tag{5}$$

基于邻域大小的主题度量方法适宜于处理长文本数据，而社会网络中包含了大量短文本消息，例如微博中的单条消息最长不超过 140 个字。仅仅通过文本的邻域来度量相似性过于简单，因此本文中加入了用户属性。社会网络中的用户相似性 $S_U(i,j)$，包括用户的微博总数 N_i 和 N_j，2 个用户之间的转发微博数 $R_{i,j}$、评论微博数 $C_{i,j}$，关注关系 FR（互相关注时 FR=2，单向关注时 FR=1）。

$$S_U(i,j) = \frac{C_{i,j} + R_{i,j}}{N_i + N_j} FR \tag{6}$$

因此，加入用户属性后，话题相似性度量通过式（7）计算，其中 β 为调节参数，一般 β=1。

$$S_{i,j} = \frac{\left|\Gamma(i) \cap \Gamma(j)\right|}{\left|\Gamma(i) \cup \Gamma(j)\right|} + \beta S_U(i,j) \tag{7}$$

在社区结构划分领域中，经典的基于模块度的方法中相似性矩阵中的元素定义为

$$S_{i,j}^{Q} = \boldsymbol{A}_{i,j} - \frac{k_i k_j}{2m} \tag{8}$$

其中，$\boldsymbol{A}_{i,j}$ 为对应的关系矩阵中的元素，其中 k_i, k_j 分别为 i、j 对应的度。基于模块度的方法具有解析度的限制、无法满足大规模数据处理等问题[7,24]，文献[24]中提出了一种新的面向社区结构划分的局部相似性度量方法，本文将其推广到有向带权网络的相似性度量为

$$S_{i,j} = \frac{\sum\limits_{V_e \in St(i) \cap St(j)} \frac{1}{D_e}}{\sqrt{\sum\limits_{V_e \in St(i)} \frac{1}{D_e}} \sqrt{\sum\limits_{V_e \in St(j)} \frac{1}{D_e}}} \tag{9}$$

其中，$St(i)$ 代表所有与 i 有连边的节点集合，D_e 的计算公式为

$$D_e = \sum_{V_e \in St(i)} W_{i,V_e} \tag{10}$$

本文面向微博用户行为分析，采用离散时间来分析用户之间的交互行为，构建离散的马尔科夫链。时间区间 Time 内，用户 i 主动对 j 发起 $C_{i,j}^{\text{Time}}$ 和 $R_{i,j}^{\text{Time}}$ 次评论和转发操作，产生的交互行为作用强度 $S_{i,j}$，因此可以在式(6)的基础上定义用户 i 对 j 在 t 时间内的作用强度。

$$S_{i,j} = \frac{C_{i,j}^{\text{Time}} + R_{i,j}^{\text{Time}}}{N_i + N_j} FR \tag{11}$$

综上所述，相似性度量方法很多，需要根据分析任务的不同及数据集的特征选择合适的相似性度量方法，本节给出了多种相似性度量方法，特别对社区结构划分、话题分析、用户行为分析的局部相似性度量方法进行了详细描述。针对社会网络中的其他数据分析，只需要给出适当的相似性度量标准，即可扩展到其他数据分析任务中。

2.3.2 扩散矩阵的构造

对于一个数据集合 X 中的任意 2 个元素 i、j 的扩散距离定义为 $d(i,j)$，对于 $\forall i, j \in X$，满足 $d(i,j) \geqslant 0$，当且仅当 i 和 j 相同时，$d(i,j)$=0。在本文中可以将 2.2 节定义的相似性度量 $S_{i,j}$ 看成扩散算子，进而计算扩散距离。根据扩散距离计算对角矩阵 $\boldsymbol{D}$，其对角元素为 $D(i)$

$$D(i) = \sum_{j=1}^{n} S_{i,j} \tag{12}$$

由于基于局部相似性度量得到的矩阵为非规范化矩阵，因此将其规范化处理，进而得到扩散矩阵 $\boldsymbol{T}$

$$\boldsymbol{T} = \boldsymbol{D}^{-\frac{1}{2}} S \boldsymbol{D}^{-\frac{1}{2}} \tag{13}$$

3 基于扩散小波的多尺度分析

社会网络在不同的尺度下表现出不同的几何结构特征，通过构造扩散小波分析，在多尺度降维的同时分析出数据的结构特征。

3.1 扩散小波的构造

在图 G 上构造基于局部相似性的扩散矩阵 $\boldsymbol{T}$，其对应的关系矩阵 $\boldsymbol{A}$，则图 G 上的拉普拉斯矩阵为 $(\boldsymbol{I}$-$\boldsymbol{T})$，因此可以考虑图 G 中的扩散过程。通过对 $\boldsymbol{T}$ 进行局部正交化操作，可以得到其压缩表示形式 $\boldsymbol{T}^2$，然后对 $\boldsymbol{T}^2$ 进行局部正交化操作，得到下一个尺度对应的压缩形式。对于尺度 j-1 时，其压缩表现形式为 $\boldsymbol{T}^{2^{j-1}}$，经过局部正交化过程可以获得压缩形式 $\boldsymbol{T}^{2^j}$，并且满足矩阵秩 Rank($\boldsymbol{T}^{2^j}$)<Rank($\boldsymbol{T}^{2^{j-1}}$)，因此能够达到降维的目的。每一个尺度计算都包括三步：减采样过程、正交化、算子压缩[14]。扩散小波的构造过程见算法 1。

算法 1 扩散小波的构造算法

$$\{\boldsymbol{\Phi}_j, \boldsymbol{\Psi}_j, T_j\} = \text{DW}(\boldsymbol{T}, \boldsymbol{\Phi}_0, J, \varepsilon, \eta)$$

输入　$\boldsymbol{T}$：扩散算子；$\boldsymbol{\Phi}_0$：初始化基矩阵，单位向量；ε：精度阈值；η：QR 分解阈值参数；J：最大计算次数。

输出　$\boldsymbol{\Phi}_j$：尺度 j 对应的缩放函数；$\boldsymbol{\Psi}_j$：尺度 j 对应的小波函数；T_j：$T_j=[T^{2^j}]_{\phi_j}^{\phi_j}$，在 $\boldsymbol{\Phi}_j$ 作用下的压缩表示。

1) for j=0 to J−1 do

2) $[\boldsymbol{\Phi}_{j+1}]_{\boldsymbol{\Phi}_j},[T^{2^j}]_{\boldsymbol{\Phi}_j}^{\boldsymbol{\Phi}_{j+1}} \leftarrow \mathrm{MQR}(T_j,\varepsilon,\eta)$

3) $T_{j+1}=[T^{2^{j+1}}]_{\boldsymbol{\Phi}_{j+1}}^{\boldsymbol{\Phi}_{j+1}}=([T^{2^j}]_{\boldsymbol{\Phi}_j}^{\boldsymbol{\Phi}_{j+1}}[\boldsymbol{\Phi}_{j+1}]_{\boldsymbol{\Phi}_j})^2$

4) $[\boldsymbol{\Psi}_j]\boldsymbol{\Phi}_j \leftarrow \mathrm{MQR}(I_{\boldsymbol{\Phi}_j}-[\boldsymbol{\Phi}_{j+1}]_{\boldsymbol{\Phi}_j}[\boldsymbol{\Phi}_{j+1}]_{\boldsymbol{\Phi}_j}^T,\varepsilon,\eta)$

5) end

在图 G 描述的数据上进行多尺度分解相当于利用扩散缩放函数 $\boldsymbol{\Phi}_j$ 生成的正交基生成子空间 $v_0 \supseteq v_1 \supseteq v_2 \supseteq \cdots \supseteq v_j$。子空间 v_{j+1} 在 v_j 中的正交补空间记为 w_j，该补空间可以通过正交扩散小波生成 $\boldsymbol{\Psi}_j$。对于尺度 j，正交基 $\boldsymbol{\Phi}_j$ 可以通过正交基 $\boldsymbol{\Phi}_{j+1}$ 来表示，记为 $[\boldsymbol{\Phi}_j]_{\boldsymbol{\Phi}_{j-1}}$，$\boldsymbol{\Phi}_j=[\boldsymbol{\Phi}_j]_{\boldsymbol{\Phi}_{j-1}}$ 为尺度 j 对应的扩散缩放函数，可以通过式(14)计算。

$$\begin{aligned}[\boldsymbol{\Phi}_j]_{\boldsymbol{\Phi}_0}&=[\boldsymbol{\Phi}_j]_{\boldsymbol{\Phi}_{j-1}}[\boldsymbol{\Phi}_{j-1}]_{\boldsymbol{\Phi}_0}\\&=[\boldsymbol{\Phi}_j]_{\boldsymbol{\Phi}_{j-1}}[\boldsymbol{\Phi}_{j-1}]_{\boldsymbol{\Phi}_{j-2}}\cdots[\boldsymbol{\Phi}_1]_{\boldsymbol{\Phi}_0}[\boldsymbol{\Phi}_0]_{\boldsymbol{\Phi}_0}\end{aligned} \tag{14}$$

算法 1 中尺度 j 对应的扩散小波函数为 $\boldsymbol{\Psi}_j=[\boldsymbol{\Psi}_j]_{\boldsymbol{\Phi}_0}$，而扩散小波函数可以通过式(14)、式(15)进行计算。

$$[\boldsymbol{\Psi}_j]_{\boldsymbol{\Phi}_0}=[\boldsymbol{\Psi}_j]_{\boldsymbol{\Phi}_{j-1}}[\boldsymbol{\Phi}_{j-1}]_{\boldsymbol{\Phi}_0} \tag{15}$$

算法 1 中采用改进的施密特分解对矩阵进行正交化分解，详细过程见算法 2。对于矩阵 $\boldsymbol{T}$，其可以分解成 $\boldsymbol{T}$=$\boldsymbol{QR}$。因此，该矩阵的正交列在精度阈值 ε 约束下得到矩阵 $\boldsymbol{Q}$ 和 $\boldsymbol{R}$，$\boldsymbol{Q}$ 为正交矩阵，$\boldsymbol{R}$ 为上三角矩阵。根据矩阵不变子空间定理可以得到 $\boldsymbol{T}$=$\boldsymbol{RQ}$，因此可以得到

$$T_{j+1}=([T^{2^j}]_{\boldsymbol{\Phi}_j}^{\boldsymbol{\Phi}_{j+1}}[\boldsymbol{\Phi}_{j+1}]_{\boldsymbol{\Phi}_j})^2 \tag{16}$$

算法 2　改进的施密特正交分解算法

$\{\boldsymbol{Q},\boldsymbol{R}\}=\mathrm{MQR}(\boldsymbol{T},\varepsilon,\eta)$

输入　$\boldsymbol{T}$：分解矩阵；ε：精度；η：分解阈值参数。

输出　$\boldsymbol{Q}$：扩展基函数；$\boldsymbol{R}$：上三角分解矩阵。

1) *colns*=|$\boldsymbol{T}$|; k=0;

2) 计算 $\boldsymbol{T}$ 的列向量范式 *Norms*;

3) while(1){k=k+1;

4) [*Norms*,*srtcln*]=sort(*Norms*,“descend”);
　　//$\boldsymbol{T}$ 范式按降序排列

5) *MaxNorm*=*Norms*(1,k);

6) if(*MaxNorm*<ε)　　break; //精度限制

7) else{*Col*=*srtcln*(1,k);//选择最大列

8)　　$\boldsymbol{Q}$=$\boldsymbol{Q}\cup$ *Col*;

9)　　$\boldsymbol{T}$(*all, Col*)=*T*(*all, Col*)/*MaxNorm*;

10)　　for j=0 to *colns* do

11)　　Orthogonalize $\boldsymbol{T}$ and obtain $\boldsymbol{T}'$;

12)　　*Norms*=*norm*();}

13)　$R=Q^T(abs(Q^T)\geqslant\eta)$;

3.2　基于扩散小波的多尺度分析算法

根据多尺度分析框架，基于扩散小波的多尺度分析算法过程见算法 3 所示。算法 3 中将框架中的关系矩阵 $\boldsymbol{A}$ 作为数据输入矩阵，通过迭代调用算法 1 针对扩散矩阵 $\boldsymbol{T}$ 进行分析。扩展基中描述了对应尺度下的几何结构特征，将扩展基 EB 和描述向量 AX 结合，能够方便理解。

由于分析对象的不同，其相似性度量方法也不同，因此采用参数 *Soptions* 来确定算法中计算扩散算子的方法。针对本文的社区结构、主题、用户行为 3 种分析任务，分别采用式(7)、式(9)、式(11)进行计算。

算法 3 中步骤 1)创建数据集的描述向量，该向量描述了数据集中节点的名字属性，作为步骤 8)中特征映射的输入。在多尺度分析的同时，输出便于理解的聚类结果。通过扩散小波构造算法 1 对扩散矩阵 $\boldsymbol{T}$ 进行迭代分析。扩散小波构造的扩展基对应尺度 j 的输出，其得到的扩展基函数就描述了数据的结构特征，根据数据描述向量和扩展基得出待分析数据的多尺度结构特征。

算法 3　基于扩散小波的多尺度分析算法

$\{EB, F, AX\}$=MSA-DW($\boldsymbol{A}$,*Soptions*, J, ε, η)

输入　$\boldsymbol{A}$：描述矩阵；*Soptions*：相似性矩阵计算参数；ε：精度；η：分解阈值参数；J：最大计算次数。

输出　*EB*：对应尺度的扩展基；*F*：几何特征描述；*AX*：数据集描述。

1) *AX*=CreateDescription($\boldsymbol{A}$); //根据 $\boldsymbol{A}$ 计算数据集描述向量

2) $\boldsymbol{S}$=ComputeSimilarity($\boldsymbol{A}$，*Soptions*); //根据

Soptions 参数类型，构建相似性矩阵

3) $\boldsymbol{T}$=CreateOperator($\boldsymbol{S}$)；//构造扩散矩阵

4) for j=1 to J do{

5) $\{\boldsymbol{\Phi}_j,\boldsymbol{\Psi}_j,T_j\}$=DW ($\boldsymbol{T},\boldsymbol{\Phi}_0,J,\varepsilon,\eta$);

//利用扩散小波构造多尺度基函数

6) $\boldsymbol{T}$==Create$\boldsymbol{T}$($\boldsymbol{\Phi}_j,\boldsymbol{\Psi}_j,T_j$);

//构造尺度 j+1 对应的扩散矩阵 $\boldsymbol{T}$

7) $\boldsymbol{EB}(j)=\boldsymbol{\Phi}_j$; //扩展基描述几何特征}

8) $\boldsymbol{F}$=MapFeature(EB,AX); //构造数据特征映射描述

4 实验及分析

在不同数据集上对社区结构、话题、用户行为的多尺度分析算法进行实验分析。平台为 Intel Core I5-3470、3.2 GHz、6 G 内存，Linux 和 Matlab 2011a，可视化工具为 NodeXL。

4.1 多尺度社区结构划分

本文分别在人工和真实数据集下分析多尺度社区结构。人工数据集采用 Andrea Lancichinettid 和 Santo Fortunato 提出的社区划分测试基准算法[26]生成 128 个节点的无向带权网络，节点度为 16，该网络包括 4 个社区，该网络如图 1 所示。4 个真实网络数据集的信息如表 1 所示，其中 Karate 和 Football 2 个为经典的社区划分数据集，LiveJournal 和 Youtube 2 个数据集为文献[27]中的数据集，考虑到实验平台中内存的限制，在原始数据集中按真实社区选择特定规模的节点。

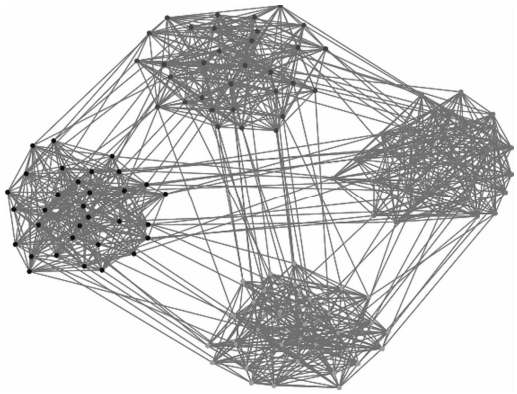
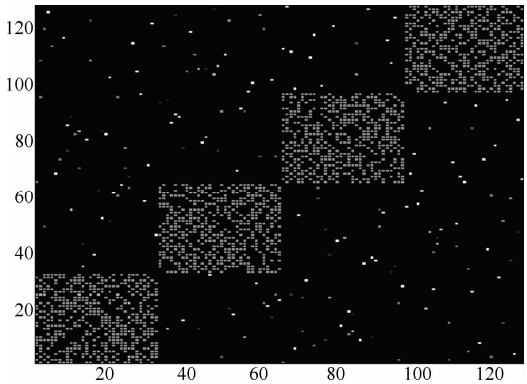


图 1 计算机生成的 128 个节点网络

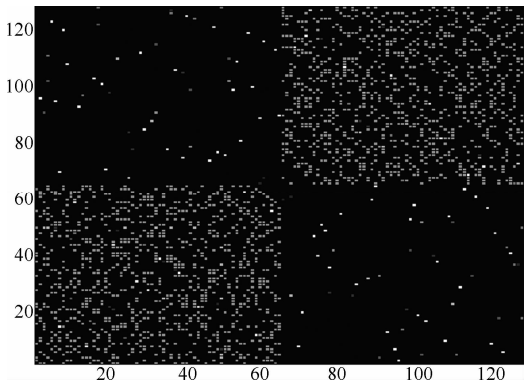
基于扩散小波的多尺度社区划分算法将图 1 所示的人工网络初始划分为 4 个社区，图 2 中(a)为 4 个社区结构对应的转换矩阵可视化示意图，而下一粗粒度的尺度被划分成 2 个社区，图 2 中(b)为 2 个社区结构对应的转换矩阵可视化示意图，最后一个尺度该网络被划分成 1 个社区，验证了扩散小波的自动收敛性。

表 1 真实网络数据集描述

数据集	节点数	边数
Karate	34	78
Football	115	613
LiveJournal	5 000	33 278
Youtube	10 000	91 040



(a) 4 个社区结构



(b) 2 个社区结构

图 2 计算机生成的 128 个节点对应的多尺度社区结构

为了验证多尺度社区划分结果的真实意义，采用真实网络数据集 Karate 进行验证。实验中设置 η 为 0.01，经过扩散小波分析后得到 7 个尺度，各个尺度对应的社区数目如表 2 所示。当尺度 j=1 时，其将每个节点划分成一个社区。当尺度 j=5 时，对应的社区划分可视化结果如图 3 所示，该网络共划分为 4 个社区，具体分割线如图 3 中虚线所示。需要特别说明的是当尺度 j=6 时，网络被划分成 2 个社区，如图 3 中实线划分所示，但是节点 3 被划分到实线左边社区，因为节点 3 在 2 个社区划分中相当于重叠节点[28]，并不影响社区划分的真实意义。当 j=7 时，将所有的节点划分成一个社区，实现自动收敛，因此无需给出参数 J。

表 2　空手道网络多尺度社区数目分布

尺度 j	社区数目
1	34
2	32
3	19
4	7
5	4
6	2
7	1

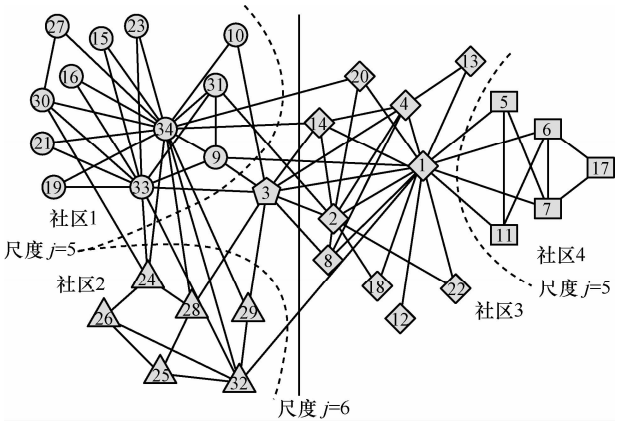


图 3　空手道数据集的多尺度社区划分

进一步选择文献[7]的链接分析算法 LINK-CLUSTER 和文献[8]中的基于全局信息标准的算法 FSM-SO、局部信息标准的算法 FSM-LFK 3 个算法进行对比分析。实验中，算法 FSM-LFK 和 FSM_SO 的尺度间隔分别为 0.05 和 1。针对表 1 中的 4 个不同规模的真实数据集进行实验，实验结果如图 4 所示。

在小规模数据集下，本文的算法与其他几种算法的互信息 NMI 值基本一致。在 LiveJournal 和 Youtube 大规模数据集中，真实网络中为分别包含 30、38 个社区，4 个算法在该尺度下划分的准确性都较高，但在社区数目大于标准值时，NMI 值都有所下降。不同数据规模下的对比结果表明，算法 MSA_DW 整体优于其他算法。

4.2　多尺度主题识别

本实验从新浪微博中抽取：“闯红灯扣分”、“丰田汽车事件”、“美国总统大选电视辩论”、“诺贝尔”、“我是特种兵 2 利刃出鞘”、“中国好声音决赛”、“烟花大会”7 个话题。实验中只考虑文本长度超过 10 个字的微博，共收集 7 000 个用户和 13 022 条微博。考虑到词频太低的词对话题影响不大，因此过滤掉了词频次数小于 20 的词。分词时专门考虑了像“中国好声音”等专有名词，处理后得到 1 107 个词。

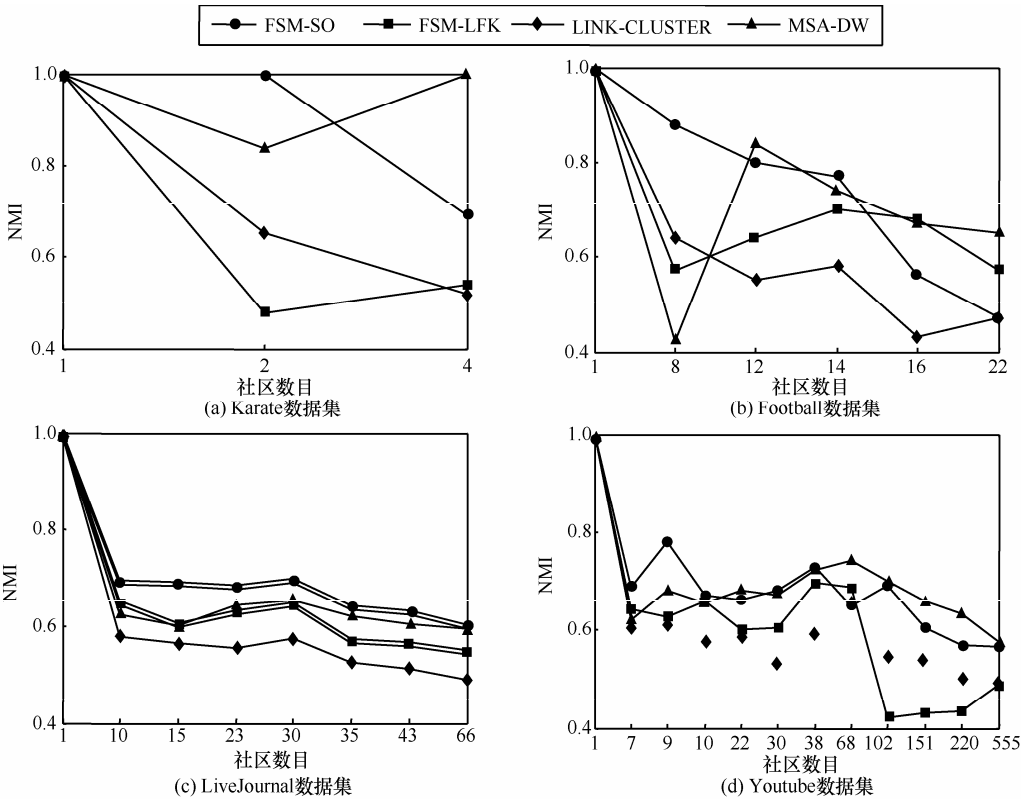


图 4　在不同数据集下的社区划分结果对比

经过扩散小波分析后得到 7 个尺度，各个尺度对应的话题数目如表 3 所示。对于尺度 j=6 时，其得到 7 个话题，这与已知话题数一致，当尺度 j=5 时，其细分为 18 个话题。表 4 中显示了尺度 j=6 对应的“丰田汽车事件”、“美国总统大选电视辩论”、“诺贝尔奖”、“中国好声音总决赛”4 个话题在尺度 j=5 时分裂成 11 个话题对应的描述词。话题“丰田汽车事件”分裂成 3 个子话题，子话题 1 描述描述丰田召回事件，子话题 2 描述丰田召回的原因，子话题 3 描述丰田休假停产。话题“诺贝尔奖”分裂成 2 个子话题，分别描述了莫言获得诺贝尔文学奖、格登.约翰获得诺贝尔医学奖 2 个子话题。因此，经过扩散小波的多尺度分析，能够在不同粒度下分析话题，并且在不同粒度下的话题都具有实际意义。

表 3　微博文本尺度与话题数目对应

尺度 j	话题数目
1	1 107
2	1 024
3	564
4	65
5	18
6	7
7	1

表 4　尺度 j=5 时 4 个话题对应的话题描述词

话题描述（j=6）	话题（j=5）	
	编号	话题描述词
“丰田汽车事件”	1	汽车 丰田 中国 销量 召回
	2	召回 全球 车窗 电动 故障
	3	生产 班次 置评 放假 停产
“美国总统大选电视辩论”	1	总统 辩论 大选 美国 奥巴马
	2	英语 字幕 视频 罗姆尼 口语
	3	钞票 餐馆 大款 文汇报 小吃店
“诺贝尔奖”	1	生理 红细胞 编程 格登.约翰 医学
	2	莫言 文学奖 诺贝尔 获奖 作家
“中国好声音总决赛”	1	中国好声音 卫视 吴莫愁 梁博 决赛
	2	参与 投票 发起 表态 心目
	3	大少爷 螺丝灯 天地-蝈蝈蝈 财经 中国好声音

进一步分析本文提出的增加用户属性的多尺度分析算法的精度，选取 j=4、5、6、7 的 4 个尺度对应的聚类结果进行精度对比实验，本文算法以 U_DWT 表示。由图 5 可知，本文的算法 U_DWT 的聚类结果比 LDA[12]和 DWT[15]话题模型的精度高，这主要得益于本文在相似性度量时加入了用户属性特征。

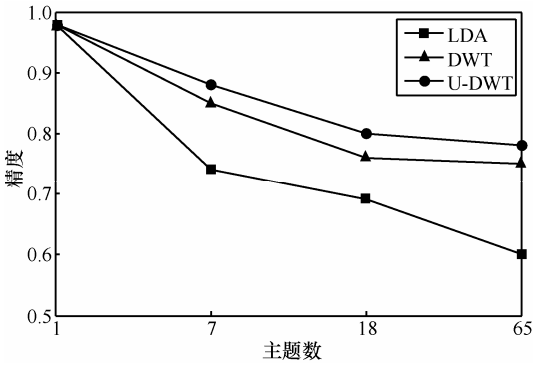


图 5　主题精度对比结果

4.3　多尺度用户行为分析

本实验分析新浪微博网络中的用户交互行为，通过新浪 API 共计抽取了 1 000 个用户，分析用户之间 10 天的交互行为，本实验中考虑转发、评论 2 种交互行为，共计 39 568 次交互。

实验中得到 4 个尺度，在尺度 j=3 和 4 时，其对应的群体聚集数都为 14 个，j=3 时的群体可视化结果如图 6 所示。但是对比分析尺度 j=3 和 4 下的微观变化，对第 10 个群体进行放大，可见在 t 时刻其对应的群体领袖为用户 12，但是对于 t+1 时刻，其对应的群体领袖为用户 222。在微博网络中，随着时间的变化，用户之间的交互产生意见领袖，但是针对特定的话题，意见领袖是动态变化的。因此，通过扩散小波的多时间尺度分析，能够得到细粒度的用户交互过程分析。

5　结束语

本文针对社会网络数据中包含的同质性关系提出了一个基于扩散小波的多尺度分析框架 MSA-DW。该框架在降维分析的同时，也将数据的几何结构特征描述出来。该框架具有很好的扩展性，只需提出适当的相似性度量标准，即可在统一的框架下实现多种同质性关系的多尺度分析。在人工和实际数据实验表明，在统一的框架下能够有效地对社会网络中的社区、话题、用户行为等进行多尺度分析。

在处理大规模数据集时，仍需要进一步提高算法的处理性能及研究尺度参数的选择。另外本文中

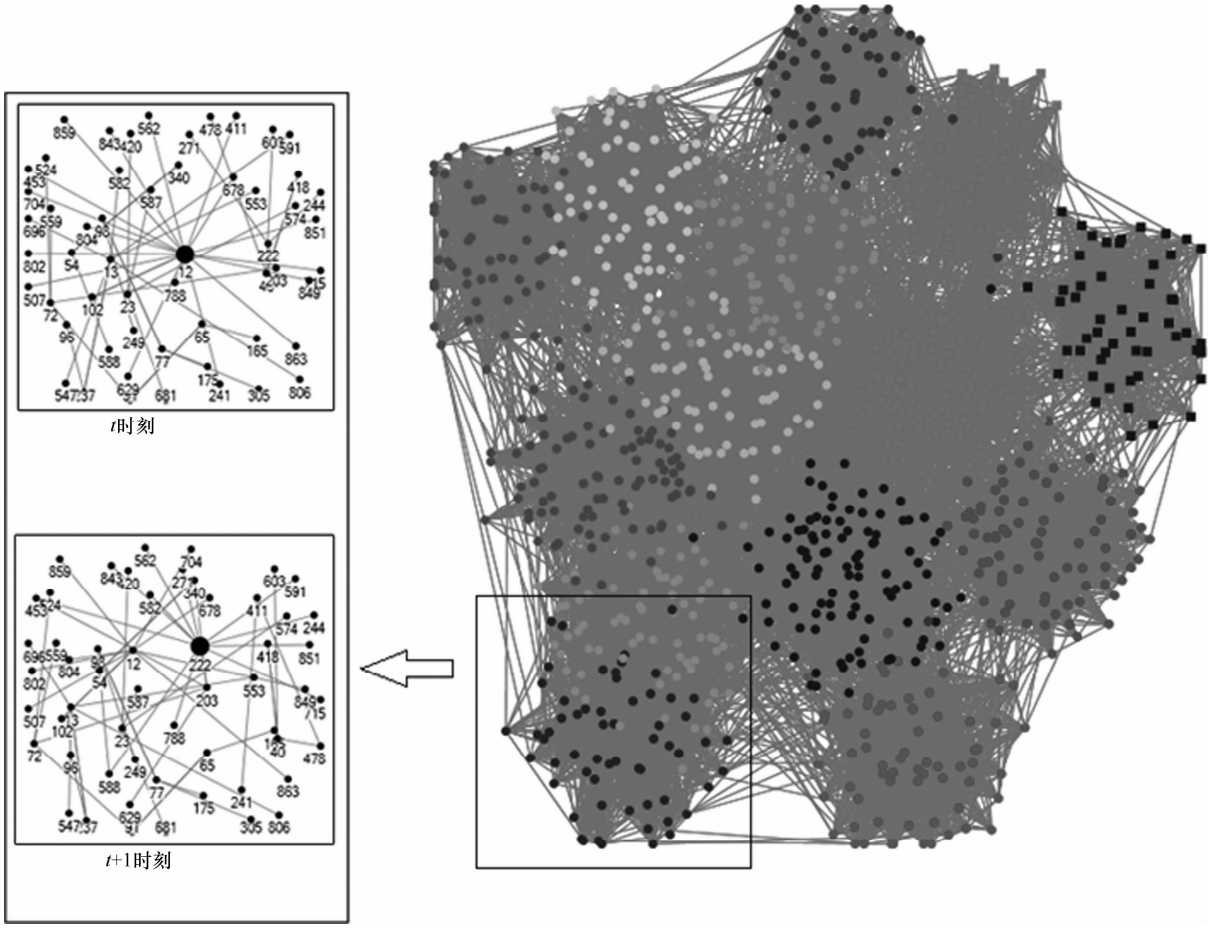


图 6　多时间尺度下的用户行为分析结果

只考虑了同质关系，需要将该框架扩展到异质关系的多尺度分析中。

参考文献：

[1] MILLER M, LYNN S L, JAMES M C. Birds of a feather: homophily in social networks[J]. Annual Review of Sociology, 2011, 27(2001): 415-444.

[2] SANTO F. Community detection in graphs[J]. Physics Reports, 2010, 486(2010):75–174.

[3] NEWMAN M E J, GIRVAN M. Finding and evaluating community structure in networks[J]. Phys Rev E, 2004,69(2) :026113.

[4] NEWMAN M E J. Modularity and community structure in networks[J]. Proceeding of the National Academy Sciences, 2006, 103(23): 8577-8582.

[5] FORTUNTO S, BARTHELEMY M. Resolution limit in community detection[J]. Proceeding of the National Academy Sciences,2007, 104(1):36-41.

[6] PETER J M, THOMAS R, KEVIN M, *et al*. Community structure in time-dependent, multiscale, and multiplex networks[J]. Science, 2010,328(5980): 876-878.

[7] AHN Y Y, BAGROW J P, LEHMANN S. Link communities reveal multiscale complexity in networks[J]. Nature,2010, 466:761-764.

[8] MARTELOT E L, HANKIN C. Fast multi-scale detection of relevant communities in large-scale networks[J]. The Computer Journal, 2013, 56(9):1136-1150.

[9] CHEN X Q, SHEN H W. Uncovering the community structure associated with the diffusion dynamics on networks[J]. Journal of Statistical Mechanics: Theory and Experiment, 2010, (2010): P04024.

[10] SHEN H W, CHEN X Q, WANG Y Z, *et al*. A dimensionality reduction framework for detection of multiscale structure in heterogeneous networks[J]. Journal of Computer Science and Technology, 2012, 27(2):341-357.

[11] LU L Y, ZHOU T. Link prediction in complex network: a survey[J]. Physical A,2011, 390(2011):1150-1170.

[12] DAVID M B, ANDREW Y N, MICHAEL I J. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3:993–1022.

[13] ZHAO W X, JIANG J, WENG J S, *et al*. Comparing twitter and traditional media using topic models[A]. The 33rd European Conference on Information Retrieval[C]. 2011.338-349.

[14] HONG L J, DAVISON B D. Empirical study of topic modeling in twitter[A]. Proceedings of the First Workshop on Social Media Analytics[C]. 2010. 80-88.

[15] WANG C, MAHADEVAN S. Multiscale analysis of document corpora based on diffusion models[A]. Proceedings of the Twenty-First International Joint Conference on Artificial Intelligence[C]. 2009. 1592-1597.

[16] GYARMATI L, TRINH T A. Measuring user behavior in online social networks[J]. Network, 2010, 24(6): 26-31.

[17] FLAVIO C, RAVI K, PRABHAKAR R, *et al*. Are Web users really markovian?[A]. Proceedings of the 21st International Conference World Wide Web[C]. 2012.609-618.

[18] WALTER W, REZA R, MAURO M. Research on social networks: time to face the real challenges[J]. ACM SIGMETRICS Performance Evaluation Review, 2009,37(3):49-54.

[19] HERVE A, LYNNE J W. Principal component analysis[J]. WIRE Computational Statistics, 2010,2(4):433-459.

[20] TENENBAUM J B, SILVA V, LANGFORD J C. A global geometric framework for nonlinear dimensionality reduction[J]. Science, 2000, 290:2319-2323.

[21] MIKHAIL B, PARTHA N. Laplacian eigenmaps for dimensionality reduction and data representation[J]. Neural Computation, 2003, 15: 1373-1396.

[22] RONALD R C, MAURO M. Diffusion wavelets[J]. Applied and Computational Harmonic Analysis, 2006, 21: 53-94.

[23] WANG C, MAHADEVAN S. Multiscale dimensionality reduction based on diffusion wavelets[EB/OL]. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.151.2341.

[24] 刘旭，易东云. 基于局部相似性的复杂网络社区发现方法[J].自动化学报,2011,37(12):1520-1529.
LIU X, YI D Y. Complex network community detection by local similarity[J]. Acta Automatica Sinica, 2011, 37(12): 1520-1529.

[25] XIANG R J, NEVILLE J, ROGATI M. Modeling relationship strength in online social networks[A]. Proceedings of the 19th International Conference on World Wide Web[C]. 2010. 981–990.

[26] LANCICHINETTI A, FORTUNATO S. Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities[J]. Physical Review E, 2009,80(1): 016118.

[27] JAEWON Y, LESKOVEC J. Defining and evaluating network communities based on ground-truth[A]. Proceedings of the 2012 IEEE 12th International Conference on Data Mining[C]. 2012.745-754.

[28] LI H J, ZHANG J H, LIU Z P, *et al*. Identifying overlapping communities in social networks using multi-scale local information expansion[J]. The European Physical Journal B, 2012, 85: 190.

作者简介：

申国伟（1986-），男，湖南邵阳人，哈尔滨工程大学博士生，主要研究方向为数据挖掘、信息安全。

杨武（1974-），男，辽宁宽甸人，博士，哈尔滨工程大学教授、博士生导师，主要研究方向为信息安全、数据挖掘、网联网安全。

黄文廷（1982-）,男，江苏徐州人，硕士，国家计算机网络与信息安全管理中心工程师，主要研究方向为信息安全、社会计算、物联网安全。

王巍（1974-），男，黑龙江哈尔滨人，博士，哈尔滨工程大学副教授，主要研究方向为数据挖掘、网络安全。

丁丽（1978-），女，安徽淮北人，硕士，国家计算机网络与信息安全管理中心高级工程师，主要研究方向为信息安全、数据挖掘。

于淼（1987-），男，黑龙江牡丹江人，哈尔滨工程大学博士生，主要研究方向为数据挖掘、社会计算。