

A Community Discovery Algorithm Based on Utility

Depin Yang¹, Lihua Zhou², Chao Cheng³, Kezhen Long⁴

Department of Computer Science and Engineering, School of Information Science and Engineering, Yunnan University, Kunming
Email: 393278091@qq.com

Received: Dec. 15th, 2013; revised: Dec. 29th, 2013; accepted: Jan. 5th, 2014

Copyright © 2014 Depin Yang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. In accordance of the Creative Commons Attribution License all Copyrights © 2014 are reserved for Hans and the owner of the intellectual property Depin Yang et al. All Copyright © 2014 are guarded by law and by Hans as a guardian.

Abstract: Based on the graph theory, this paper proposes a community discovery algorithm based on utility. The method considers not only the contact frequency but also the importance of the links between community members. The paper defines the concept of utility to describe nodes similarity and implements the CDBU algorithm. The algorithm proposed effectively avoids the abuse of the traditional community discovery method based on the contact frequency, which neglects the important degree of contact. Finally, on the real-world dataset, we verify the rationality and validity of the algorithm proposed in this paper.

Keywords: Social Network; Community Discovery; Utility; K-Means

一种基于效用的社区发现算法

杨德品¹, 周丽华², 程超³, 龙克珍⁴

云南大学信息学院计算机科学与工程系, 昆明
Email: 393278091@qq.com

收稿日期: 2013 年 12 月 15 日; 修回日期: 2013 年 12 月 29 日; 录用日期: 2014 年 1 月 5 日

摘要: 本文从图论思想出发, 提出了一种基于效用的社区发现算法, 该方法既考虑了社区成员联系的频繁度又考虑了联系的重要度。本文定义了效用的概念, 通过效用来描述节点相似度, 并实现了基于效用的社区发现(Community Discovery Based on Utility, 简称为 CDBU)算法, 该算法有效地避免了传统的基于联系频繁度的社区发现方法忽略了联系重要度的弊端。最后, 本文在真实数据集上进行了实验, 验证了所提出算法的合理性和有效性。

关键词: 社会网络; 社区发现; 效用; K 均值

1. 引言

随着微博、论坛、社交网站等的迅速普及, 社会网络越来越成为人们生活中不可或缺的一部分。社会网络分析已经成为一个越来越重要的研究课题^[1-3], 不仅在传统的社会学领域里受到大家的热捧, 在疾病传播、论文合作撰写、电子商务以及电子推荐系统等领域也受到了广泛关注。近年来, 通过对社会网络的分析研究, 发现社会网络中普遍存在着小世界^[1]、无标度^[2,4]以及聚集^[5]等基本统计特性。此外, 社会网络的另一个主要特征就是其所呈现出的社区结构。社区是由一系列彼此相似而与网络中其他节点不相似的节点组成的, 具有社区内部的连接比较紧密, 社区之间的连接比较稀疏的特点。社区发现技术是数据挖掘和探查性分析中的一个极为重要的技术, 它在社会网络分析、机器学习、生物学、犯罪学^[5-10]等相关领域有着十分广泛的应用。

在真实网络中,成员之间往往存在着相互联系。传统的社区发现算法要么只考虑网络成员之间的邻接关系和拓扑结构,并不考虑成员交互的频繁度,即无论发生了几次联系都记为“1”,例如 Newman^[6]等提出的基于模块度的社区发现方法,淦文燕^[11]等提出的基于拓扑势的网络社区发现方法。要么只考虑联系的频繁度,并把频繁度值作为边权进行社区发现,例如吴文涛^[12]等提出的基于权重信息挖掘社会网络中隐含社团的方法,孔兵^[13]提出的基于交互度的带权图社区发现方法。这些方法都忽略了联系的重要度这个重要因素,事实上,成员间发生的每一次联系它的重要程度往往是不一样的,对于某个体而言有的联系意义很小,甚至完全没有意义。基于这种考虑,本文受效用理论启发,对社区结构进行了重新定义,提出了一种基于效用的社区发现算法,该方法既考虑了社区成员联系的频繁度又考虑了联系的重要度,使得挖掘出来的社区更具有实际意义。因为在实际网络社区中,用户关心的往往不是与哪些成员经常有联系,而是关心哪些成员对于该用户来说是最具价值的,即用户希望得到的是对其有重要价值的成员所组成的社区,本文把这样的社区定义为用户的效用社区。

2. 相关工作

效用(Utility),是经济学中最常用的概念之一。效用值是以决策者的当前现状为基础的精神感受值,如果把某一方案导致的结果记为 X , 则该方案的结果对于该决策者的效用值记为 $U(X)$ 。效用是一个主观概念,通常由用户自己或者领域专家决定,可以表示经济收入、成本风险、审美价值和个人偏好等,具体依领域实际而定。例如在销售领域中,卖一件牛奶和卖一台电脑的效用是不一样的,甚至卖多件牛奶的效用都远小于卖一台电脑的效用。这是因为效用包含内部效用和外部效用,总效用由它们共同决定^[14]。在此,内部效用是某商品在一次交易中被卖出的数量,而外部效用是该商品的单位利润。

效用挖掘作为一种新的挖掘技术,研究起步比较晚。效用挖掘的正式引入是在2003年 IEEE 第三次数据挖掘国际会议上,由 Raymond Chan 等首次提出的,随后一些研究者陆续参与到该领域的研究中来,例如 Longbing Cao 等提出的挖掘高效用序列模式的算法^[14],熊学栋提出的基于效用模式的挖掘算法^[15],这些研究基本上都是基于效用的传统关联规则的数据挖掘研究。

当前,有关社会网络社区发现研究已取得大量的成果,但基于效用理论的社区挖掘方法目前还没有相关文献资料和成果。在社区挖掘中,影响社区划分的因素有联系的频繁度及其对应的重要度,相当于于效用挖掘中的内部效用和外部效用。因此,可以把效用理论引入到社区挖掘问题中,这是一个全新的研究课题,具有重要的研究意义,这也正是本文研究的出发点和落脚点。

3. 基于效用的社区发现算法

3.1. 问题描述

真实世界中,网络的不同节点对之间通常是带权的,比如在一个科研论文合著撰写的社会网络中,每一个节点代表一位合著者,而两个节点之间的一条边代表了一次论文合著撰写记录。在传统的研究中通常只考虑两个人的合著的次数,却忽略了每次合著的重要度研究,只是简单地认为合著越频繁就越重要,显然,这对于网络的划分没有更多的实际指导意义。因为在实际网络中,往往存在这样的情况:虽然合著者 A 与 B 之间合著次数很多,但可能合著的重要程度不是很高;相反,虽然 A 与 C 合著次数很少,但合著的的重要程度却很高。如图 1 所示。

这是具有 4 个节点的完全图,有序数对 (a,b) 代表边上的权重, a 表示合著发表在某期刊上的论文数, b 表示该期刊的影响因子。从图中可以看出, A 和 B, C 和 D 的合著次数远大于其它节点之间的合作次数,如果不考虑期刊的影响因子,社区发现算法将把 AB 划为一个社区, CD 划为另一个社区。但考虑影响因子后划分的结果可能完全不同。基于以上考虑,本文提出了一种新的基于效用理论的算法^[13,16]。

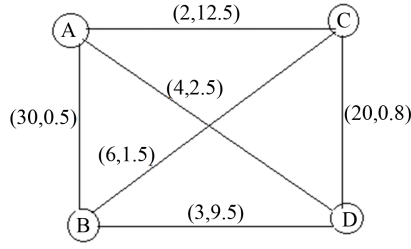


Figure 1. Scientific research cooperation network
图 1. 科研合作网

3.2. 相关定义

本文认为，如果社区个体之间存在某种联系，则认为他们之间是有关联的。为了定量地描述社区个体之间的联系，引入了一个带权无向图，定义如下：

定义 3.1 带权无向图 $G(V, E, W)$ ，其中， V 是节点集合， $E = \{(u, v) | u \in V, v \in V\}$ 是边集合， W 是边上所带的权重。

在此，怎样描述和确定权重 W 是关键，将直接影响社区划分的结果。

定义 3.2 效用函数 f 是一个二元函数 $f(q, w)$ ，记为 $U = f(q, w)$ ， U 表示效用值，且有：

$$U = \left(\sum_{i=1}^n q_i^k w_i \right)^{\frac{1}{k}} \quad (1)$$

式(1)中 q_i 表示指标值(本文中为频繁度值)， k 表示幂次，取值一般为正整数或负整数， w_i 表示权重。通常情况下为了方便研究，一般对 q_i 和 w_i 做标准化处理。

当 $k = 1$ 时，式(1)变为加权算术平均合成模型：

$$U = \sum_{i=1}^n q_i w_i = q_1 w_1 + q_2 w_2 + \cdots + q_n w_n$$

当 $k = -1$ 时，式 (1)变为加权调和平均合成模型：

$$U = \frac{1}{\frac{1}{q_1} w_1 + \frac{1}{q_2} w_2 + \cdots + \frac{1}{q_n} w_n}$$

当 $k = 2$ 时，式(1)变为加权平方平均合成模型：

$$U = \sqrt{q_1^2 w_1 + q_2^2 w_2 + \cdots + q_n^2 w_n}$$

理论上， k 的值可以为任意正整数和负整数甚至小数，但考虑到人们的思维习惯和实际评价中的需要，一般以上三种函数形式就足以满足评价需要了。

定义 3.3 U_{ij} 表示节点 i 与节点 j 之间联系的效用值，用来度量节点 i 与节点 j 之间的联系价值度。

根据上述公式，本文对图 1 的每对节点之间的效用值进行计算，结果如下： $U_{AB} = 15$ ， $U_{CD} = 16$ ， $U_{AC} = 25$ ， $U_{BD} = 28.5$ ， $U_{AD} = 10$ ， $U_{BC} = 9$ ，直观地，我们把 AC 划分在一个社区，BD 划分在一个社区，这与仅考虑合著的频繁度划分的结果是完全不同的，这也正是本文研究的价值所在。因此，通过计算节点之间的效用值构造效用矩阵，然后调用 MATLAB 中的 eig 函数求解该效用矩阵的最大 k 个特征值及其对应的 k 个特征向量就可以得到软社区指标。另外，求最大特征值即求最大效用，也就是求最大联系价值度的过程。

基于上述思想，本文把效用值即联系价值度作为社区之间相似度的度量指标，提出了基于效用的社区发现算法。为计算方便，对 U 进行了归一化处理，归一化公式如下：

$$U_{\max} = \max_{i=1 \sim n, j=1 \sim n} \{u_{ij}\}$$

$$U_{ij} = \frac{u_{ij}}{U_{\max}}, U_{\max} > 0$$

$U_{\max}$ 是所有 u_{ij} 中的最大值, u_{ij} 是归一化后的节点效用值。由公式可以看出, 若 $u_{ij} = 0$, 则 $U_{ij} = 0$, 若 $u_{ij} = U_{\max}$, 则 $U_{ij} = 1$ 。 U_{ij} 越趋近 1, 表示联系价值度越高; U_{ij} 越趋近 0, 表示联系价值度越低。即 $0 \leq U_{ij} \leq 1$ 。

3.3. 基于效用的社区发现算法

基于效用的社区发现算法(Community Discovery Based on Utility, 简称为 CDBU)的主要思想是: 将数据集中的每个对象看作是图的节点 V , 将边 E 上的效用值 U (即联系价值度)作为节点间的相似度来度量, 这样我们就得到一个基于相似度的无向带权图 $G(V, E, W)$, 于是聚类问题就转化为图的划分问题。其主要步骤为:

- 1) 计算节点间的效用值 U_{ij} ;
- 2) 构建网络对象集的效用矩阵 U ;
- 3) 通过 eig 函数计算效用矩阵 U 的最大 k 个特征值与特征向量, 构建特征向量空间;
- 4) 调用 K 均值聚类算法对特征向量空间中的特征向量进行聚类。

算法 CDBU 伪代码描述:

输入:

$G(V, E, W)$: 无向带权图

k : 社区个数

输出: 社区集合 S

- 1) $n = |V|$ /* 节点数 */
- 2) FOR $i = 1$ TO n DO /* 计算节点间的效用值 */
- FOR $j = i + 1$ TO n DO /* 不考虑节点与自己的效用值 */

$$U = \left(\sum_{i=1}^n q_i^k w_i \right)^{\frac{1}{k}}$$

IF $u_{ij} > 0$ THEN /* 归一化效用 */

$$U_{\max} = \max_{i=1 \sim n, j=1 \sim n} \{u_{ij}\}$$

$$U_{ij} = \frac{u_{ij}}{U_{\max}}$$

END IF

END j

END i

- 3) $[V, D] = \text{eig}(U)$ /* 求效用矩阵 U 的特征值 D 及特征向量 V */
- 4) 调用 K-means 算法 /* 对特征向量聚类划分 */
- 5) 输出社区集合 S

4. 实验结果与分析

在真实数据集上进行实验, 验证本文所提算法的合理性和有效性。实验环境为因特尔酷睿 i3 处理器, 内存 2G, 32 位 windows 7 操作系统, 主要编程语言为 C 语言, 由于涉及大量的矩阵操作, 故借助了 MATLAB 编程

工具。

4.1. 科研合作数据

该数据集来源于知网检索,是机械设计及理论领域的一个论文合作撰写数据,数据集一共包含 38 个合作者。论文合作期刊有清华大学教育研究、清华大学学报、计算机工程与应用和导弹与航天运载技术等共 16 个期刊。部分原始数据整理后基本样式如表 1 所示。

为使合作网络数据更加直观和便于理解,通过合作者之间的论文合作撰写关系,本文构建了他们之间的合作关系网络,其基本网络结构如图 2 所示。

Table 1. Scientific research collaboration data

表 1. 科研合作数据

作者	期刊	时间	被引	下载	影响因子
申永胜; 汤晓瑛; 姚华; 方嘉秋	高等工程教育学院	1999-07-25	15	143	2.492
阎绍泽; 申永胜; 陈洪彬	清华大学学报	2003-02-28	25	471	0.517
孟宪飞; 申永胜; 郝智秀; 王晓箐; 汤晓瑛	计算机应用	2003-05-28	21	163	0.465
索双富; 季林红; 阎绍泽; 王人成; 申永胜	实验技术与管理	2004-02-29	8	97	1.340
孟宪飞; 申永胜; 汤晓瑛; 王晓箐	计算机工程与应用	2002-08-01	11	67	0.423
阎绍泽; 陈洪彬; 李海阳	导弹与航天运载技术	2002-06-10	1	189	0.456

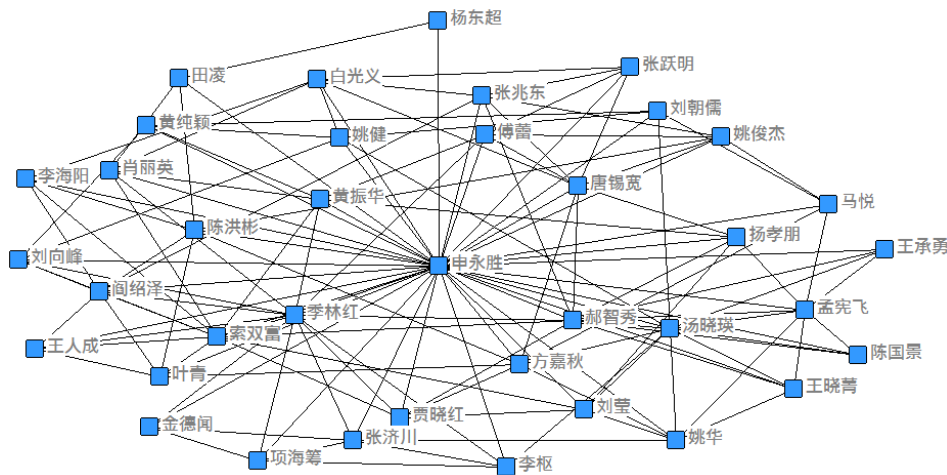


Figure 2. Scientific research cooperation network

图 2. 科研合作网络

4.1.1. 实验效果对比

通过合作网络,本文分别构建了基于链接(仅考虑邻接关系,即无论发生几次联系都记为“1”）、基于频繁度和基于效用的合作矩阵,并通过 eig 函数分别求解其对应的特征向量,再调用 K 均值算法对特征向量进行划分。以社区数目 $K=4$ 为例,三种方法的社区划分结果如图 3~5 所示。

从上面三个图中可以看出,基于链接和基于频繁度的社区划分结果差异不是特别大,但基于效用划分的结果与前两种方法得到的结果差异较大。这是由于基于效用的方法,考虑了合作者之间的多重影响因素,得到的结果是按合作的综合影响力即合作价值度来划分的。由图 5 可知,在整个网络中,申永胜、阎绍泽、季林红和索双富之间的合作综合影响力较大,也就是他们之间的合作效用或者合作的价值度比较大。类似地,郝智秀、

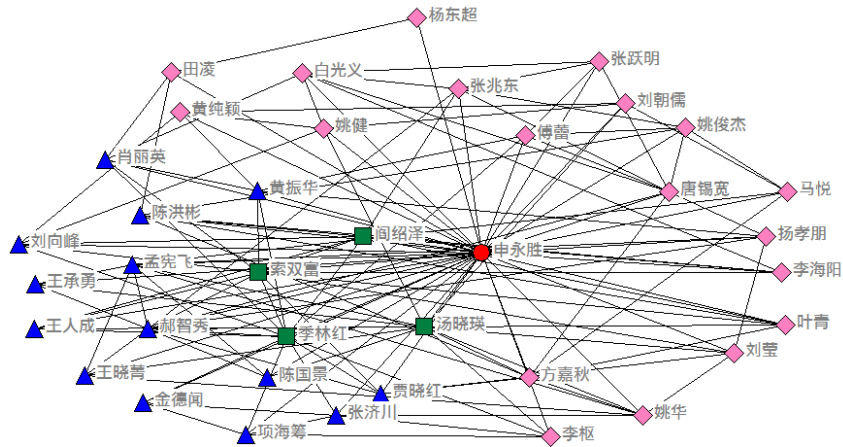


Figure 3. Community divided based on link
图 3. 基于链接的社区划分

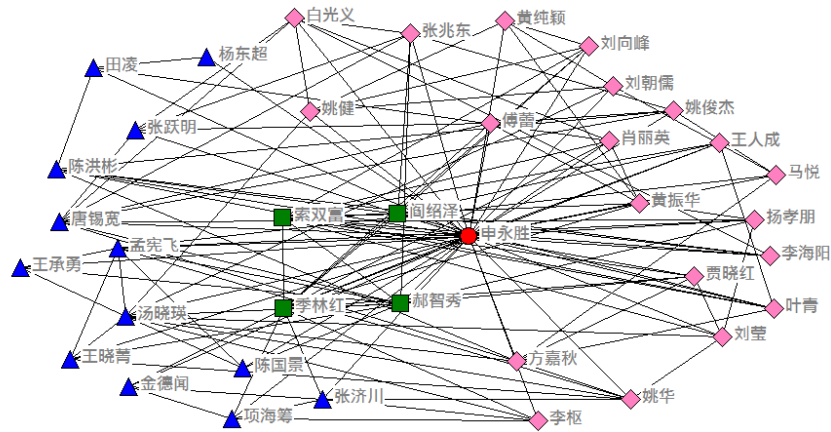


Figure 4. Community divided based on frequent
图 4. 基于频繁度的社区划分

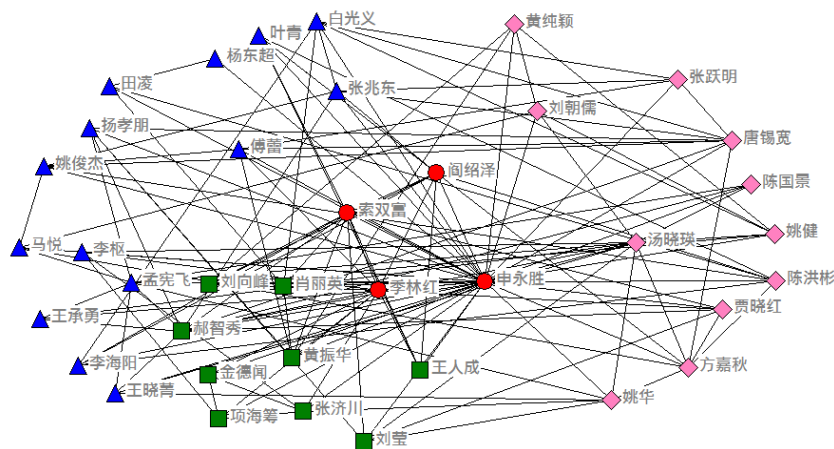


Figure 5. Community divided based on utility
图 5. 基于效用的社区划分

刘莹、王人成、黄振华、肖丽英、刘向峰、金德闻、项海筹和张济川之间的合作效用较大，并依次按图标颜色

和形状形成四个效用社区。

实验表明，基于效用的方法能够挖掘出社会网络中价值度最大的隐含社区，对于人们的来说，这比传统的社区划分方法更有实际指导意义，因为在实际中人们关心的往往不是划分了哪些社区，而是希望找出对人们最有价值的社区，即上述的效用社区。

4.1.2. 实验性能对比

首先给出的是算法的时间性能。本文在随机数据上对不同社区数目和不同节点数目的算法执行时间进行了比较，结果如图 6 所示。

从图中可以看出，当节点数目小于 10,000 个，即中小网络规模时，随着节点数目的增加，算法的执行时间增加并不明显，也就是节点数目对算法性能的影响不大。当节点数目大于 10,000 个，即大网络规模时，算法的执行时间显著增加，此时节点数目对算法性能影响较大。此外，随着社区数目的增加，算法的执行时间也逐渐增加，说明社区划分数目对算法性能有一定的影响。

现在来看社区划分的精度。在样本相似性度量的基础上，聚类分析还需要误差准则函数来把真正属于同一类的样本聚成一个簇。误差准则函数用来评价聚类结果的质量，本文选用的是误差平方和准则函数 E ，该函数的值在算法的每次迭代过程中逐渐减小，直到最终收敛至一个固定的值，此时算法迭代结束^[17,18]。函数定义如下：

$$E = \sum_{i=1}^k \sum_{p \in c_i} |p - m_i|^2$$

其中 E 是数据库中所有对象的平方误差的总和， p 是空间中的点，即给定的数据对象， m_i 是簇 c_i 的平均值。该准则函数目的是使划分的簇尽可能地紧凑和独立， E 值越大，说明误差越大，聚类结果越不好。反之，若 E 值越小，则误差越小，聚类结果就越好。因此，我们的目标是寻求使 E 最小的聚类结果，也就是在误差平方和准则下的最优结果。

通过运用上述误差准则函数，本文对各方法在不同社区划分数目上的误差进行了实验，实验结果如图 7。

从图中可以看出，无论是基于链接、基于频繁度还是基于效用，误差都是随着社区数目 K 的增大而递减的，

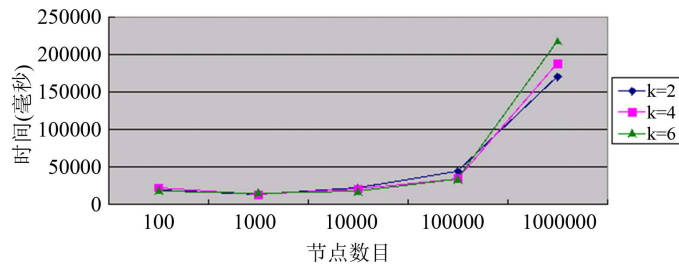


Figure 6. Algorithm time performance
图 6. 算法时间性能图

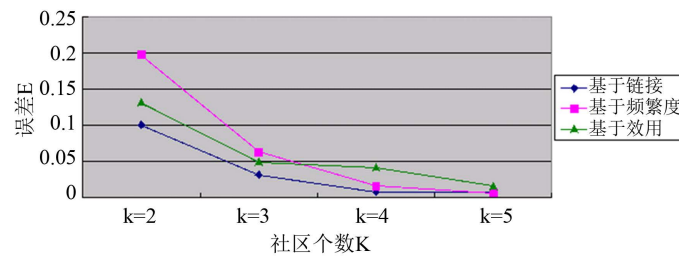


Figure 7. Algorithm error comparison
图 7. 算法误差对比

即划分的社区数目越多结果越精确。此外，本文提出的基于效用的算法虽然在误差控制上称不上是最优的，但误差也已经相当小，并且随着 K 的增大性能越发明显，说明了本文提出的算法是有效的。

5. 结论

由于社会网络挖掘问题的复杂性、挑战性以及在数据挖掘等各领域的广泛应用，使得研究社会网络及其挖掘算法具有十分重要的意义。本文提出了一种基于效用的社区发现算法，该方法既考虑了社区成员联系的频繁度又考虑了联系的重要度。实验比较了基于链接、基于频繁度和基于效用的社区划分误差，表明基于效用的方法具有优异的误差控制，说明了算法的有效性。实验也对不同社会网络规模、不同社区划分数目的算法执行时间进行了比较，表明算法在中小网络规模上具有良好的性能。下一步工作将在本文基础上，将算法扩展到多关系社会网络结构上，研究多关系社会网络中的社区发现问题。

6. 致谢

首先要感谢我的导师周丽华教授！周老师专业功底深厚、治学严谨，对问题有独到的见解，有幸跟随周老师学习，是我整个学习生涯中的宝贵经历。在论文撰写以及实验的整个过程中，周老师都极具耐心地予以指导和帮助，并在最终阶段对论文进行了字斟句酌地细致修改。此外，在生活和工作方面，周老师也给予我很大的支持和帮助。在此，谨向周老师致以深深的敬意和最衷心的感谢！

此外，还要感谢实验室程超、龙克珍等同学给予的支持和帮助，在我学习和论文研究的过程中，你们热心的帮助和指点，使我顺利完成了研究工作。

基金项目

国家自然科学基金 No. 61262069。

参考文献 (References)

- [1] Watts, D.J. and Strogatz, S.H. (1998) Collective dynamics of small-world networks. *Nature*, **393**, 440-442.
- [2] Barabási, A.L. and Albert, R. (1999) Emergence of scaling in random networks. *Science*, **286**, 509-512.
- [3] Albert, R., Barabási, A.L. and Jeong, H. (2000) The Internet's Achilles heel: Error and attack tolerance of complex networks. *Nature*, **406**, 378-382.
- [4] Barabási, A.L., Albert, R., Jeong, H. and Bianconi, G. (2000) Power-law distribution of the World Wide Web. *Science*, **287**, 2115a.
- [5] Girvan, M. and Newman, M.E.J. (2002) Community structure in social and biological networks. *Proceedings of the National Academy of Science*, **9**, 7821-7826.
- [6] Newman, M.E.J. and Girvan, M. (2004) Finding and evaluating community structure in networks. *Physical Review E*, **69**, Article ID: 026113.
- [7] Kleinberg, J.M. (1999) Authoritative sources in a hyperlinked environment. *Journal of the ACM*, **46**, 604-632.
- [8] Palla, G., Derenyi, I., Farkas, I. and Vicsek, T. (2005) Uncovering the overlapping community structures of complex networks in nature and society. *Nature*, **435**, 814-818.
- [9] Barabási, A.L. and Oltvai, Z.N. (2004) Network biology: Understanding the cell's functional organization. *Nature Reviews Genetics*, **11**, 101-113.
- [10] Yang, B., Liu, D.Y., Liu, J.M., Jin, D. and Ma, H.B. (2009) Complex network clustering algorithms. *Journal of Software*, **20**, 54-66 (in Chinese with English Abstract). <http://www.jos.org.cn/1000-9825/3464.htm>
- [11] 涂文燕, 赫南, 李德毅, 王建民 (2009) 一种基于拓扑势的网络社区发现方法. *软件学报*, **8**, 2241-2254.
- [12] 吴文涛, 肖仰华, 何震瀛, 汪卫, 余韬 (2009) 基于权重信息挖掘社会网络中的隐含社团. *计算机研究与发展*, ISSN 1000-1239/CN 11-1777/TP46, 540-546.
- [13] 孔兵 (2012) 基于连接度量的社区发现研究. 博士学位论文, 云南大学, 昆明.

- [14] Yin, J.F., Zheng, Z.G. and Cao, L.B. (2012) USpan: An efficient algorithm for mining high utility sequential patterns. KDD'12, Beijing, 12-16 August 2012, 660-668.
- [15] 熊学栋 (2007) 数据库中关联规则及效用模式挖掘算法研究. 硕士学位论文, 湘潭大学, 湘潭.
- [16] 林旺群, 卢风顺, 丁兆云, 吴泉源, 周斌, 贾焰 (2012) 基于带权图的层次化社区并行计算方法. *软件学报*, **6**, 1517-1530.
- [17] 王实 (2002) 数据挖掘中的聚类算法. *计算机科学*, **4**, 42-45.
- [18] 孙孝萍 (2002) 基于聚类分析的数据挖掘算法研究. 硕士生论文.