

基于链路动态变化的产业网络预测模型研究

王 斌^{1,2}, 王文平², 费为银¹

(1. 安徽工程大学数理学院, 安徽 芜湖 241000;

2. 东南大学经济管理学院, 江苏 南京 211189)

摘要: 根据历年的产业网络, 提出了一种预测任意两个产业将来发生链接可能性的新方法. 利用网络分析方法, 以中国的投入产出表为研究对象, 构建了基于链路动态变化的产业网络预测模型, 对比分析了该模型与相似性链路预测模型的预测精度. 研究表明, 历年产业网络的拓扑结构非常接近. 依据正交试验设计, 在参数选择合理的前提下, 本文提出的预测模型具有理想的预测精度. 相比于相似性链路预测算法, 基于链路动态变化的产业网络预测算法体现了过往信息对产业网络链路预测的重要性, 具有更好的预测结果.

关键词: 产业网络; 预测模型; 预测精度; 正交试验设计

中图分类号: TP273

文献标识码: A

文章编号: 1000-5781(2018)06-0721-11

doi: 10.13383/j.cnki.jse.2018.06.001

Study of the prediction model of industrial network based on the dynamical links

Wang Bin^{1,2}, Wang Wenping², Fei Wei Yin¹

(1. School of Mathematics and Physics, Anhui Polytechnic University, Wuhu 241000, China;

2. School of Economics and Management, Southeast University, Nanjing 211189, China)

Abstract: Based on the industrial networks over the past years, this paper proposes a new methodology that can predict the likelihood of an existing link between any nodes in the future. With the aid of network analysis, a prediction model of the industrial network is built for Chinese input-output tables. Its prediction accuracy is compared with that of the similarity link prediction model. The results show that the industrial network topology structures are very close over the years. The link prediction model has good prediction accuracy as long as the parameters are selected reasonably according to orthogonal experimental design. Compared with the similarity link prediction algorithm, the proposed methodology reflects the importance of past information to the industrial network link prediction, and has better prediction results.

Key words: industrial network; prediction model; prediction accuracy; orthogonal experimental design

1 引 言

经济投入产出系统能够被构建成产业网络, 节点表示系统中的各个产业或企业, 链路表示相互之间的物质、副产品和能量的循环流动. 网络分析方法能够探讨产业网络的结构和性质^[1], 能够测量网络中各个节点

收稿日期: 2016-09-16; 修订日期: 2017-04-10.

基金项目: 安徽高校自然科学研究重点资助项目(KJ2018A0109); 安徽工程大学引进人才科研启动基金资助项目(2018YQQ-022); 国家自然科学基金面上资助项目(71873002); 国家社会科学基金重大招标资助项目(12&ZD207).

之间的交流,而且能够解释相互之间的影响和信息流动^[2-4],因而它是分析经济投入产出系统的一个有力工具.目前,产业网络的构建主要基于投入产出表所反映的信息,并据此研究产业网络结构特征及其演化,为产业结构的优化、以及产业的转型升级提供合理的政策建议.如文献[5-10]从整体的角度,以某个年份的投入产出表为基础,研究产业网络的无标度性、幂律分布和最大关联树等基本特征.文献[11-15]从个体的角度,以几个年份的投入产出表为基础,研究产业网络内的主导产业,如何在国家产业结构调整和优化升级过程中发挥各自的优势和作用,以及探讨产业网络的演化,分析集群网络之间的互动及部门角色的变动问题.可见,投入产出表对于研究产业网络是多么的重要.但从时间维度看,基于投入产出表之上的产业网络,只是研究了某些年份的产业网络结构特征,而没能克服借助静态、历史的数据研究产业网络的局限.再从现实情况看,囿于编制投入产出表需要消耗大量的人力和物力,中国的投入产出表只是每五年编制一次,且表的公布年份距编表年份至少也会有2年~3年.较长的编表周期导致投入产出表注定是一种历史数据,那么分析以该表为基础的产业网络,对于经济结构变化较大的转型经济体来说,数年的滞后可能会影响研究结果的正确性.因此,及时、准确地预测产业网络具有重要的意义.

若能克服投入产出表跨度大,不能及时、准确反映产业网络结构变化的缺点,使得产业网络的预测变得便捷可行,则能依据若干个年份的投入产出表,对产业网络进行及时、动态的分析.事实上,随着时间的演化,产业网络中原有的链路可能消失或产生新的链路,从而表现出了较强的动态性和复杂性,处理如此问题选择社会网络分析法(SNA)十分适宜^[16],链路预测又是网络分析法中的重要任务之一.链路预测是指,依据已知的网络结构等信息,预测网络中尚未链接的两个节点之间将来产生链接可能性的大小,该理论已经被广泛应用于各种网络分析当中^[17-19].因此,链路预测是尽可能的推断出两个节点之间链接存在的可能性,具有重要的理论和实际意义.

在真实的网络中,相比于用实验结果去推断两个节点之间是否有相互作用关系,利用链路预测去预估节点将来的行为,或者识别两个节点之间未来的链接,成本更为低廉.例如,在新陈代谢网络与蛋白质相互作用网络中,需要通过大量实验结果推断节点之间是否存在相互作用关系^[20],高额的实验成本自然不可避免.降低实验成本,并且不失准确地推断结论的理想方法,是针对这些网络的结构特性、设计出一套足够精确的链路预测算法,然后实验在预测结果的指导下进行.已经有许多学者研究了链路预测的问题,这些研究主要是基于当前网络的结构性质,预测任意两个节点之间将来发生链接的概率,且研究的科学领域众多.例如,Goldberg等^[21]利用蛋白质网络的局部集聚性,预测了缺失的链接;Manning等^[22]构建了一个信息检索网络,节点代表关键词或分类文件,通过预测两者之间的关系,区分了未经辨识的文件.图书推荐网络是链路预测的另一个事例,Chen等^[23]在用户和书目之间建立了许多图论测度,以期向用户做图书推荐.此系统是一个用户-图书的二部网络,链路表示用户对某种图书的偏好.链路预测也可以被用于分析演化网络,例如,Zhou等^[24]利用链路预测理论分析了互联网将来的形状;Juszczyszyn等^[25]利用马尔科夫链构建了一个链路预测模型,分析了大学邮箱网络的子图结构随时间的演化情况.解决链路预测问题一个最简单的方法是所谓的得分算法,其中基于相似性的得分预测算法能够得到很好的预测结果,并且网络的拓扑结构性质能够帮助选择合适的相似性指标^[26,27].随后,学者们又考虑了加权网络上的链路预测问题,发现含权指标的得分预测效果要好于无权的预测方法^[28].这些研究都充分体现了得分预测算法的简洁性,并且得分预测方法很容易和其他的方法相结合.例如,Aaron等^[29]应用层级结构知识预测了丢失的链接,其中层级结构可以解释许多网络所表现出的右偏态度分布、高集聚系数和最短路径长度性质.Chungmok等^[30]从网络度分布的角度,用数学规划方法预测了网络将来的结构,其中,预测问题被转化为整数规划问题,这样做的目的是以便最大化链路预测得分总和.Chen等^[31]构建了一个快速的相似性链路预测算法,并以真实的世界网络进行了实证分析,结果表明新的算法比其他算法速度更快,而且精度更高.总之,学者们提出了诸多方法去解决链路预测问题,这些方法主要以测量节点之间相似性为基础.其中目前应用最广泛的测量方法、即相似性预测算法有结构等价指标 CN、资源分配指标 RA、约旦系数指标 JC 和阿达米克-亚达指标 AA^[32-34].

但是,以上的研究较少涉及到产业网络的链路预测问题,并且随着时间的演化,产业网络中原有的链路可能消失或产生新的链路.相似性链路预测算法,虽然可以根据当前的产业网络,预测任意两个产业将来发生链接的概率.但是,产业网络的拓扑结构随时间而变化,产生了大量涉及节点和链路的动态信息,若能在链路预测的算法当中体现出这些有用信息,则预测的精度会有进一步的提高.为了检验这个设想,本文以中

国 2005 年、2007 年、2010 年和 2012 年的四个投入产出表为研究对象, 首先构建四个有权重的产业网络, 然后基于权重网络提出一种新的链路预测算法, 建立基于链路动态变化的产业网络预测得分算法, 识别两个产业之间建立链接的可能性大小. 与相似性预测算法相比, 这种新的算法不仅考虑两个节点之间的相似性程度, 受到共同邻居节点的影响(本文用控制变量 α 表示), 而且考虑在前期网络和当前网络中, 链路权重的变化程度(本文用变化率 r 表示). 在考虑了变化程度的基础上, 又进一步考虑了链路权重变化的方向, 即权重的增减情况(本文分别用三个参数 δ, η, θ 表示). 此外, 为了测试本文提出的预测算法性能优劣, 又进一步引入了目前应用最广泛的相似性链路预测算法作为对比, 结果显示本文提出的链路预测模型的预测精度更高. 因此, 在预测产业网络的链路时, 不仅要考虑当前产业网络的链路情况, 还要充分考虑产业网络中链路的动态变化信息, 这样得出的结果才会更加准确可靠.

2 基于链路动态变化的产业网络预测模型

产业网络具有动态的演化结构, 这些结构随时间的演化而变化, 新的节点不断产生, 新的链路不断形成, 以及其上的权重不断改变, 使得产业网络结构具有动态性. 针对产业网络中链路动态变化的特性, 并考虑到两个节点邻居的得分情况, 本文提出的预测算法, 在权重的特定比例上定义预测得分的增加或减少.

2.1 产业网络模型的构建

本文旨在研究产业网络中的链路预测方法, 因此, 首先讨论如何构建产业网络, 而投入产出表是构建产业网络的基础, 且能准确表达各个产业部门在生产与分配领域的经济联系. 本文分析的投入产出表, 是以直接消耗系数为元素所构成的矩阵. 记 $\mathbf{A} = (a_{ij})_{n \times n}$ 为直接消耗系数矩阵, 其中 $a_{ij}, i, j = 1, 2, \dots, n$ 为直接消耗系数, 且 $a_{ij} \in [0, 1]$. 直接消耗系数反映了产业部门生产一个单位的总产品所需要消耗其他部门产品的比例, 它是一个无量纲的数值, 恒在 $[0, 1]$ 区间变化, 因此不会发生随着经济的增长其绝对值随之增加的情况, 并且本文是根据历年的投入产出情况, 预测任意两个产业部门之间未来的链接情况, 而直接消耗系数既可以反映两个部门之间有无联系, 又可以反映出联系的紧密程度, 其比值越接近于 1, 联系越紧密; 反之越稀疏. 以直接消耗系数为链路, 产业为节点构建的产业网络考虑的是任意两个产业部门之间的链接情况, 与方向无关, 因此本文研究无向网络. 同时, 考虑到链接的重要性, 借鉴刘刚等^[6]的处理方法, 仅将两链接量作均值处理, 再将所有产业对链接量的均值设为阈值, 阈值以上的值定为有效链接.

2.2 产业网络的结构指标

1) 平均度

网络的平均度定义为节点度数中心性的平均值, 节点 i 的度数中心性表示与其相连边的个数 k_i , 其表达式为

$$k_i = \sum_{j \in \Gamma} e_{ij}, \quad (1)$$

其中 Γ 表示与节点 i 直接相连的节点组成的集合, 亦称为节点 i 的邻居节点集. $e_{ij} = 1$ 时表示节点 i 与 j 存在连边, 否则 $e_{ij} = 0$. 度数中心性 k_i 越大, 表示产业部门 i 在产业网络中与产业部门的联系就越多, 在相应的投入产出表中作用就越重要, 与其它产业部门进行的物质、副产品和能量的交互往来就越多.

2) 密度

网络的密度 ρ 反映网络节点间联系的紧密程度, 表达式定义为

$$\rho = 2L / (N(N-1)), \quad (2)$$

其中 L 为网络中实际存在的有效关联数, N 为网络中所有产业部门的个数. ρ 越大, 表示在相应的投入产出表中, 产业部门间的联系越紧密.

3) 簇系数

在网络中, 网络的簇系数是所有节点簇系数的均值, 而节点的簇系数定义为

$$C_i = 2e_i / (k_i(k_i - 1)), \quad (3)$$

其中 e_i 表示节点 i 的邻居节点之间实际存在的边的个数, k_i 表示节点 i 的度数中心性.

网络的簇系数表示的内涵是, 你的朋友圈或熟人圈中的每个人都是相互认识的. 事实上, 因为你的朋友大部分是你的同事、同学和邻居, 所以他们互相认识的概率自然应该很大.

4) 平均最短距离

平均最短距离 d 是网络的一个重要结构指标, 网络中所有节点之间的平均最短距离定义为

$$d = \frac{2}{N(N-1)} \sum_{i \geq j} d_{ij}, \quad (4)$$

其中 N 为网络中所有节点的个数, d_{ij} 为连接节点 i 和 j 最短路径上的边的个数.

网络中的搜索、路由等相关算法的高效实现皆与平均最短距离紧密相关. 在相应的投入产出表中, 平均最短距离越小, 表示任意两个产业部门之间的物质、副产品和能量的流动就越便捷.

2.3 演化情形

本文定义产业网络中链路动态变化, 为网络中任意两个节点之间物质流从一种状态到随后另一种状态的变化情况. 观察历年的产业网络可以发现, 节点之间的权重有衰减、保持不变和增加三种情形. 由于节点对 (u, v) 之间的权重 $\omega(u, v)$ 往往随着时间的变化而变化, 因此可用 $\omega(u, v, t), t \in [0, \infty)$ 表示网络中节点对 (u, v) 之间的权重 $\omega(u, v)$ 是时间 t 的函数. 在产业网络中, 任意一对节点之间的权重增减数值各不相同, 为了便于定义随后的预测得分, 本文引进变化率 $r (0 < r < 1)$, 基于任一 t_1 时刻的产业网络中任意一对节点 (u, v) 之间的权重 $\omega(u, v, t_1) (\neq 0)$, 定义三个集合 $E_1 = [0, (1-r)\omega(u, v, t_1)], E_2 = [(1-r)\omega(u, v, t_1), (1+r)\omega(u, v, t_1)], E_3 = [(1+r)\omega(u, v, t_1), \infty)$, 显然 $E_1 \cap E_2 \cap E_3 = \phi$, 而 $E_1 \cup E_2 \cup E_3 = \Omega$. 然后, 视权重 $\omega(u, v, t), t > t_1$ 与三个集合的隶属关系, 把权重的演化情况分别分为衰减、保持不变和增加三种类型.

1) 衰减

当产业网络从 t_1 时刻的状态演化到 t 时刻的状态时, 若节点对 (u, v) 之间的权重 $\omega(u, v, t)$ 较权重 $\omega(u, v, t_1)$ 为减少, 且 $\omega(u, v, t) \in E_1$ 时, 定义此时的衰减函数为

$$D(u, v, t) = \begin{cases} \int_{t_1}^t \delta \omega(u, v, t) dt, & \omega(u, v, t) \in E_1 \\ 0, & \text{其它,} \end{cases} \quad (5)$$

其中 δ 为负数, 表示权重在衰减.

由于节点对之间的权重从 t_1 时刻的状态演化到 t 时刻的状态时, 其变化过程中权重是连续变化的, 因此衰减函数可用积分表示.

2) 保持不变

当产业网络从 t_1 时刻的状态演化到 t 时刻的状态时, 若节点对 (u, v) 之间的权重 $\omega(u, v, t)$ 较权重 $\omega(u, v, t_1)$ 变化不大, 即 $\omega(u, v, t) \in E_2$ 时, 定义此时的不变函数为

$$C(u, v, t) = \begin{cases} \int_{t_1}^t \eta \omega(u, v, t) dt, & \omega(u, v, t) \in E_2 \\ 0, & \text{其它,} \end{cases} \quad (6)$$

其中 η 为非负数, 表示权重的变化细微, 可以忽略不计.

3) 增加

当产业网络从 t_1 时刻的状态演化到 t 时刻的状态时, 若节点对 (u, v) 之间的权重 $\omega(u, v, t)$ 较权重 $\omega(u, v, t_1)$ 为增加, 且 $\omega(u, v, t) \in E_3$ 时, 定义此时的增加函数为

$$I(u, v, t) = \begin{cases} \int_{t_1}^t \theta \omega(u, v, t) dt, & \omega(u, v, t) \in E_3 \\ 0, & \text{其它,} \end{cases} \quad (7)$$

其中 θ 为非负数, 表示权重在增加.

由以上的定义可知, 参数 θ, δ 和 η 的关系应为 $\theta > \eta > \delta$, 三者的数值将在评估预测算法的性能时确定, 选择预测精度最高的参数值。

2.4 基于链路动态变化的产业网络预测得分

与相似性预测算法不同, 本文提出的算法不仅考虑两个节点之间的相似性程度受共同邻居节点的影响, 而且考虑前期网络和当前网络中, 链路权重的变化程度, 在考虑了变化程度的基础上, 又进一步考虑链路权重变化的方向, 即权重的增减情况, 来定义节点对 (u, v) 在区间 $[t_1, t_2], t_2 > t_1$ 上的预测得分 $\text{score}(u, v)$ 为

$$\text{score}(u, v) = \int_{t_1}^{t_2} (P(u, v, t) + \alpha S(u, v, t)) dt, \quad (8)$$

其中 $P(u, v, t) = D(u, v, t)I_{E_1} + C(u, v, t)I_{E_2} + I(u, v, t)I_{E_3}$, $S(u, v, t) = \sum_{y \in \Gamma(u) \cap \Gamma(v)} [P(u, y, t) + P(y, v, t)]$, $\Gamma(u)$ 表示节点 u 的邻居节点的集合, $I_{E_i}, i = 1, 2, 3$ 是 E_i 的示性函数。

$P(u, v, t)$ 计算了产业网络中节点对 (u, v) 从 t_1 时刻的状态演化到 t 时刻的状态时的得分。假定节点 y 是节点 u, v 共同的邻居节点, 令 $S(u, v, t)$ 表示分别与节点 u, v 相邻的节点对 (u, y) 与 (y, v) , 从 t_1 时刻的状态演化到 t 时刻的状态时的所有得分之和。参数 α 是控制变量, 表示邻居节点影响到节点 u 和 v 之间关系的程度。 $\text{score}(u, v)$ 就是基于链路动态变化的产业网络预测模型的得分计算公式。特别, 在 $\omega(u, v, t_1) = 0$ 的情形下, 若 $\omega(u, v, t) = 0$, 则设 $P(u, v, t) = 0$; 若 $\omega(u, v, t) \neq 0$, 则设 $P(u, v, t) = \theta$ 。由此构建的式(8)包含了两个节点间的相似性程度受共同邻居节点的影响, 节点间的链路权重变化程度, 以及变化的方向等动态信息, 其中利用三个连续区间 E_1, E_2 和 E_3 , 把投入产出表每五年更新一次, 这种离散的跳跃, 分别归类到这三个区间, 据此构建连续变化性模型, 来预测当前产业网络的链接情况。

2.5 模型的预测算法精度实验设计及敏感性分析

2.4 节提出的基于链路动态变化的产业网络预测模型, 既考虑了两个节点之间的相似性程度受共同邻居节点的影响, 又考虑了在前期网络和当前网络中, 链路权重的变化程度以及变化的方向。那么该模型的算法精度如何, 本文将引入相关指标作进一步分析。目前, 共有三种衡量链路预测算法精度的指标, 分别为 AUC(area under the receiver operation characteristic curve), Precision 和 Ranking score^[35]。它们对预测精度衡量的侧重点不同。由于本文以整个产业网络为研究对象, 因此使用 AUC 指标从整体上来衡量算法的精度, 并且产业网络是小规模的网络, 因此本文在计算 AUC 时, 采用逐项遍历方法, 即每次从网络中选取一条边进行测试, 余下的边作为训练集, 然后测试这条边, 得到一个相应的预测精度。最后遍历网络中的每条边, 计算平均值, 作为整个网络的预测精度。同时, 在基于链路动态变化的产业网络预测算法式(8)中, 为了得到算法精度最优的参数值, 本文将采用正交试验设计^[36], 及其统计方法确定所需的参数, 并分析数值模拟结果对相关参数的敏感性。

3 基于链路动态变化的产业网络预测模型实证分析

本文以中国投入产出表为例, 选出 2005 年、2007 年、2010 年和 2012 年的中国投入产出直接消耗系数表作为研究对象, 以直接消耗系数为链路, 产业部门为节点构建产业网络。由于本文提出的预测算法, 主要考虑产业网络中链路动态变化的情况, 即在产业网络中保持节点个数不变, 考察随着时间的演化, 网络中任意两个节点之间链路的断开或链接情况, 然后据此预测下一期的产业网络。而在中国历年的投入产出表中, 产业部门的名称和数目(网络中的节点)并不是完全一致。因此, 本文采用文献[37]的处理方法, 只对前后不统一的产业部门, 即 2005 年的旅游业, 2007 年、2010 年、2012 年的水利、环境和公共设施管理业, 对这两个产业部门进行合并与整理, 而其他的产业部门不会改动, 因此在删除这两个产业部门后, 对其余产业部门之间、产生的新链路或消失的链路影响甚微。鉴于此, 把 2005 年投入产出表中旅游业所在的行与列删除, 把 2007 年、2010 年、2012 年投入产出表中的水利、环境和公共设施管理业所在的行与列删除。利用前文构建产业网络的方法, 得到四个年份的有权重的、且是无向的产业网络, 如图 1 所示, 从左到右, 从上到下分别是 2005 年、2007 年、2010 年和 2012 年的产业网络。

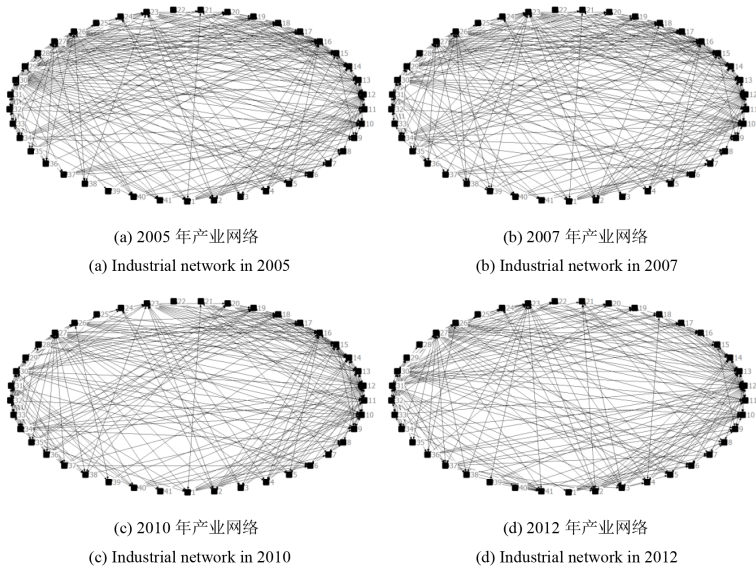


图 1 四个年份的中国产业网络

Fig. 1 China's industrial network in four years

利用式(1)~式(4), 对各个年份产业网络的结构指标进行分析, 得到的结果如表 1 所示.

表 1 各个年份产业网络的参数
Table 1 Parameters of industrial network in each year

产业网络	平均度	密度	簇系数	平均最短距离
2005 年	11.610 0	0.289 6	0.598 0	1.962 0
2007 年	11.220 0	0.280 5	0.552 0	1.968 0
2010 年	10.878 0	0.272 0	0.568 0	1.976 0
2012 年	10.146 0	0.253 7	0.555 0	2.070 0
均值	10.962 5	0.274 0	0.568 3	1.989 5
标准差	0.621 7	0.014 1	0.020 0	0.060 0

由表 1 可知, 平均度和密度逐年减小, 平均最短距离逐年增大, 而簇系数没有明显的变化规律. 结果表明物质、副产品和能量的循环流动可以到达更远的产业部门, 资源的利用更加充分, 但从数值上看, 这种变化又不太显著. 四个年份的产业网络的平均度、密度、簇系数和平均最短距离的标准差很小, 各项指标相似, 说明四个产业网络的结构非常接近, 因此可以构建新的链路预测模型.

3.1 产业网络的预测

由于我国投入产出表每五年编制一次, 中间年份再出一次延长表, 分别于 1997 年、2002 年、2007 年和 2012 年四个年份出版, 而 2000 年、2005 年和 2010 年的表格是投入产出表的延长表. 其中 2005 年以前的投入产出表距本文所预测的 2015 年投入产出关系已达十多年之久, 对于预测结果无显著的影响. 因此, 本文将用 2005 年、2007 年、2010 年和 2012 年的产业网络, 去预测下一年份即 2015 年产业网络的链接情况. 这四个年份的结构指标在上节已做了详细分析, 结果非常接近. 因此, 可以利用基于链路动态变化的产业网络预测模型中的式(8), 计算 2015 年产业网络中任意两个节点之间的得分, 再利用 2.1 节中产业网络模型的构建方法, 得到 2015 年的产业网络. 进而, 可将 2015 年产业网络的预测结果、与过去几年的真实数据相比较, 分析产业的转型升级、迁移等情况. 下面将以产业网络结构指标中的度数中心性为例, 分析产业重要性的演化情况. 利用式(1), 计算预测得到的 2015 年产业网络中各个产业的度数中心性, 再与 2005 年和 2010 年相比较, 结果如表 2 所示.

从表 2 可以看出, 批发和零售业度数中心性在 2005 年的网络中居于第九位, 2010 年网络中上升到第二位, 最后来到了 2015 年网络中的第一位, 重要性可见一斑. 事实上, 批发和零售业是各个产业部门的产品实

现价值的重要媒介,是连接商品生产和消费的主要环节.除了批发和零售业以外,化学工业度数中心性的位置也在提升,由2005年的第七位,到2010年的第四位,最后上升到2015年的第二位,可见其在当下产业网络中的重要性.度数中心性发生显著变化的还有金融业,由2005年和2010年的前十位以外进入到2015年的第四位,它在产业网络内度数中心性位置的提升,说明了其是配置社会资源和融通资金作用的源动力,是现在经济生活的命脉和媒介.批发和零售业与金融业属于第三产业,“十二五”规划结束后,第三产业的度数中心性在前十个产业中占到了六席,产业结构优化调整的效应从预测得到的产业网络中得到了充分的体现.为了得到预测方法精度,下节将给出计算预测算法精度的实验设计和敏感性分析.

表2 三个年份度数中心性排在前十的产业部门

Table 2 Three-year-degree centrality ranked in the top ten of the industry sector

降序	2005 年	2010 年	2015 年
1	电力、热力的生产和供应业	交通运输及仓储业和邮政业	批发和零售业
2	研究与试验发展业综合技术服务业	批发和零售业	化学工业
3	交通运输及仓储业和邮政业	电力、热力的生产和供应业	交通运输及仓储业和邮政业
4	租赁和商务服务业	化学工业	金融业
5	水利、环境和公共设施管理业,居民服务和其他服务业	研究与试验发展业综合技术服务业	金属冶炼及压延加工业
6	信息传输、计算机服务和软件业	租赁和商务服务业	通用、专用设备制造业
7	化学工业	电气机械及器材制造业	租赁和商务服务业
8	建筑业	通信设备、计算机及其他电子设备制造业	住宿和餐饮业
9	批发和零售业	金属制品业	非金属矿物制品业
10	金属冶炼及压延加工业	建筑业	公共管理和社会组织

3.2 产业网络预测算法精度实验设计及敏感性分析

实例数据来源于上节所构建的2005年、2007年、2010年和2012年的产业网络,依据预测精度指标AUC,通过正交试验设计和对参数的敏感性分析,给出基于链路动态变化的产业网络预测模型参数的最优设定值.

1) 正交试验设计

在基于链路动态变化的产业网络预测算法式(8)中,为了得到算法精度最优的参数值,本文将采用正交试验设计及其统计方法确定所需的参数,并分析参数的敏感性.正交试验设计(orthogonal experimental design)是多因素多水平的实验设计方法^[36],依据具体问题选择合适的正交表是使用正交法的关键.在基于链路动态变化的产业网络预测算法中,共有五个参数,依据各个参数所表示的不同意义, r 、 α 和 θ 分别选取两个水平, δ 选取四个水平, η 选取五个水平,在此基础上设计正交表.然后,依据正交表进行仿真实验,并将结果列表示意. α 的两个水平分别设为0.05和0.1,先分析 $\alpha = 0.05$ 的情况,再分析 $\alpha = 0.1$ 的情况.当 $\alpha = 0.05$ 时,利用网络的两个版本进行研究.一个是基于四个权重产业网络,当变化率 r 设为25%,另一个变化率 r 设为50%.计算不同版本下的AUC值并列表显示.当变化率 r 设为25%,且 $\alpha = 0.05, \theta = 2$ 时,通过MATLAB 7.0编程计算得到AUC值,如表3所示.当变化率 r 设为25%,且 $\alpha = 0.05, \theta = 3$ 时,AUC的值,如表4所示.

表3 变化率 r 设为25%时AUC的值($\alpha = 0.05, \theta = 2$)Table 3 The AUC values under $r = 25\%$ ($\alpha = 0.05, \theta = 2$)

$\delta \setminus \eta$	0.00	0.50	1.00	1.50	2.00
-2.00	0.771 4	0.821 5	0.859 1	0.898 6	0.923 8
-1.50	0.914 9	0.913 5	0.919 5	0.928 3	0.935 9
-1.00	0.934 5	0.935 3	0.938 0	0.942 2	0.946 0
-0.50	0.945 8	0.946 5	0.948 0	0.946 7	0.951 8

表4 变化率 r 设为25%时AUC的值($\alpha = 0.05, \theta = 3$)Table 4 The AUC values under $r = 25\%$ ($\alpha = 0.05, \theta = 3$)

$\delta \setminus \eta$	0.00	0.50	1.00	1.50	2.00
-2.00	0.795 0	0.835 4	0.874 7	0.909 5	0.930 6
-1.50	0.921 7	0.921 0	0.925 6	0.933 2	0.940 2
-1.00	0.938 0	0.938 1	0.940 2	0.943 6	0.947 5
-0.50	0.946 7	0.946 8	0.947 7	0.949 3	0.950 9

当变化率 r 设为50%,且 $\alpha = 0.05, \theta = 2$ 时,AUC的值,如表5所示;当变化率 r 设为50%,且 $\alpha = 0.05, \theta = 3$ 时,AUC的值,如表6所示.

从表3~表6可以看出,预测精度AUC最大是表5中的值0.979 3,相应的参数 δ, η, θ 取值分别为-0.5, 0.5, 2,同时 $r = 0.5$.当 $\alpha = 0.1$ 时,仍然利用产业网络的两个版本进行研究.一个是基于四个权重

产业网络, 当参数 α 变化率 r 设为 25%, 另一个变化率 r 设为 50%, 计算不同版本下的 AUC 值. 计算的方法同 $\alpha = 0.05$ 时的情况, 这里就不再赘述. 比较分析得到的结果发现, 当 α, r, δ, η 和 θ 的取值分别为 0.05、0.5、-0.5、0.5 和 2 时, 预测精度 AUC 最大.

表5 变化率 r 设为 50% 时 AUC 的值($\alpha = 0.05, \theta = 2$)

Table 5 The AUC values under $r = 50\%$ ($\alpha = 0.05, \theta = 2$)

$\delta \setminus \eta$	0.00	0.50	1.00	1.50	2.00
-2.00	0.808 5	0.906 8	0.934 1	0.960 3	0.974 0
-1.50	0.969 5	0.968 6	0.971 6	0.975 6	0.977 7
-1.00	0.977 3	0.977 4	0.978 1	0.978 8	0.978 9
-0.50	0.979 2	0.979 3	0.979 2	0.978 8	0.978 3

表6 变化率 r 设为 50% 时 AUC 的值($\alpha = 0.05, \theta = 3$)

Table 6 The AUC values under $r = 50\%$ ($\alpha = 0.05, \theta = 3$)

$\delta \setminus \eta$	0.00	0.50	1.00	1.50	2.00
-2.00	0.830 9	0.905 7	0.941 6	0.963 7	0.975 2
-1.50	0.970 3	0.968 5	0.970 6	0.974 1	0.976 8
-1.00	0.975 1	0.974 5	0.974 7	0.975 7	0.976 7
-0.50	0.975 8	0.975 1	0.974 9	0.974 9	0.975 2

2) 参数的敏感性分析

由以上分析的结果可知, 当分别以 $r = 0.5$ 和 $r = 0.25, \alpha = 0.05$ 和 $\alpha = 0.1$, 对预测精度 AUC 进行变化分析时, 得到链路预测算法中各个参数 α, r, δ, η 和 θ 的取值分别为 0.05、0.5、-0.5、0.5 和 2 时, 预测精度 AUC 最大. 为了研究数值模拟结果对这两个参数的敏感性, 在此做一个敏感性分析, 以一定区间的取值范围为参考, 考虑 r 和 α 对预测精度 AUC 的变动趋势. 不失一般性, r 的取值范围为 [0.1, 0.8], 间距设为 0.05; α 的取值范围为 [0, 1.5], 间距设为 0.05, 计算其对预测精度 AUC 的影响, 得到趋势图 2. 当 r 在 [0.1, 0.8] 范围内取值时, 得到预测精度 AUC 的极差为 0.058 3, 均值为 0.960 3 和标准差为 0.018 6; 而当 α 在 [0, 1.5] 范围内取值时, 得到预测精度 AUC 的极差为 0.161 3, 均值为 0.867 3 和标准差为 0.052 8. 可见, 预测精度对这些参数的变化不很敏感, 但 α 比 r 的敏感性要高, 说明在计算预测精度时, 优先考虑到两个节点的关系受到邻居节点的影响程度.

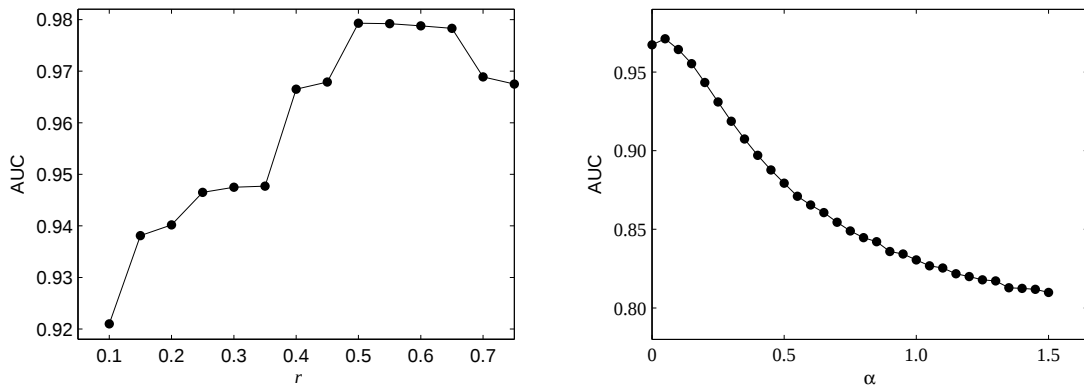


图2 AUC 的变化趋势

Fig. 2 The change trend of AUC values

3.3 预测算法精度的对比分析

为了测试本文提出的预测算法性能的优劣, 进一步引入相似性链路预测算法作为对比. 相似性链路预测算法指标有两类, 一类是基于无权网络的相似性指标, 如式(9)~式(12)所示; 另一类是基于权重网络的指标, 如式(13)~式(16)所示. 它们分别为 CN 指标(common neighbors, 又称结构等价指标)、RA 指标(resource allocation 指标, 资源分配指标)、JC 指标(Jaccard's coefficient 指标, 约旦系数指标)和 AA 指标(Adamic-Adar指标, 阿达米克-亚达指标)^[32-34]. 这八个指标的具体表示式为

$$\text{CN}(u, v) = |\Gamma(u) \cap \Gamma(v)|, \quad (9)$$

$$\text{RA}(u, v) = \sum_{z \in \Gamma(u) \cap \Gamma(v)} \frac{1}{|\Gamma(z)|}, \quad (10)$$

$$\text{JC}(u, v) = \frac{|\Gamma(u) \cap \Gamma(v)|}{|\Gamma(u) \cup \Gamma(v)|}, \quad (11)$$

$$AA(u, v) = \sum_{z \in \Gamma(u) \cap \Gamma(v)} \frac{1}{\ln(|\Gamma(z)|)}, \quad (12)$$

$$CN(u, v) = \sum_{w \in \Gamma(u) \cap \Gamma(v)} [\omega(u, w) + \omega(v, w)], \quad (13)$$

$$RA(u, v) = \sum_{z \in \Gamma(u) \cap \Gamma(v)} \frac{\omega(u, z) + \omega(v, z)}{2s_z}, \quad (14)$$

$$JC(u, v) = \sum_{w \in \Gamma(u) \cap \Gamma(v)} \frac{\omega(u, w) + \omega(v, w)}{\sum_{a \in \Gamma(u)} \omega(a, u) + \sum_{b \in \Gamma(v)} \omega(b, v)}, \quad (15)$$

$$AA(u, v) = \sum_{w \in \Gamma(u) \cap \Gamma(v)} \frac{\omega(u, w) + \omega(v, w)}{\ln(1 + \sum_{a \in \Gamma(w)} \omega(w, a))}, \quad (16)$$

其中 $\Gamma(u)$ 和 $\Gamma(v)$ 分别表示节点 u 和 v 的邻居节点的集合, $|\Gamma(z)|$ 表示节点 z 的度, s_z 表示节点 z 的强度。

在文献[32–34]中, 没有考虑节点 u, v 之间的权重 $\omega(u, v)$ 与时间的关系, 因此式(13)~式(16)中只利用 $\omega(u, v)$ 表示节点 u, v 之间的权重, 而本文提出的预测模型中考虑了节点 u, v 之间的权重 $\omega(u, v, t)$ 与时间相关, 并据此提出了基于链路动态变化的产业网络预测模型。许小可等^[38]研究发现, 如果任意两节点之间的最短距离长度大于等于 2 时, 那么可以使用以上基于共同邻居算法的八种链路预测算法。由于在 2005 年、2007 年、2010 年和 2012 年的产业网络中, 任意两节点之间的最短距离长度大于等于 2, 因此, 可以使用相似性链路预测算法作比较。

截至目前, 由于 2015 年包含 42 部门的投入产出表, 尚未对外公布, 因此无法分析 2015 年产业网络的真实情况。因此本文采用链路预测中的常用做法, 再结合本文研究的产业网络的特征, 利用 2.5 节中所介绍的预测精度指标 AUC, 去衡量本文算法与相似性链路预测算法哪个更加优越。现将相似性链路预测算法的精度 AUC, 计算得到的结果, 列于表 7。

表 7 相似性链路预测算法的精度 AUC

Table 7 The AUC values of prediction accuracy based on similarity link prediction algorithm

预测算法	权重网络				无权网络				本文算法
	AA	CN	JC	RA	AA	CN	JC	RA	
预测得分	0.781 2	0.757 1	0.551 9	0.840 8	0.779 7	0.723 8	0.426 5	0.7859	0.979 3

其中 AUC 的最大值为 0.840 8, 而基于链路动态变化的产业网络链路预测算法的精度为 0.979 3。与相似性链路预测算法的精度相比较, 显然本文算法的预测精度最大。因此在预测产业网络的链路时, 不仅要考虑产业网络当前的链接情况, 还要充分考虑产业网络的演化情况, 这样得出的结果才会更加准确可靠。

4 结束语

当考虑了产业网络的链路动态变化因素时, 通过对中国 2005 年、2007 年、2010 年和 2012 年产业网络的拓扑结构及链路权重变化的分析, 提出了一种新的产业网络预测算法模型。新的预测算法不仅考虑了产业网络中链路的动态变化信息, 而且考虑到两个节点邻居的得分情况, 在权重的特定比例上定义预测得分的增加或减少, 最后把每个变化过程的得分相加即得产业网络的链路预测得分。实证结果表明, 利用正交实验设计方法, 只要其中的参数选取合适, 则新的产业网络预测算法模型具有理想的预测精度。并且相比于相似性链路预测算法, 本文提出的基于链路动态变化的产业网络预测模型的预测精度更加理想。因此链路上的动态变化信息对预测产业网络具有重要意义。

为使产业网络的预测更加准确, 研究者们应该充分考虑过往信息对产业网络链路预测的重要性。由于数据的易得性, 本文仅仅考虑了中国产业网络的演化情况。其实, 产业网络(产业共生网络)形态各异, 既有宏

观上的产业网络,又有中观或微观上的产业网络,深入研究这些网络的链路预测很有意义,可以指导产业的升级或迁移等问题.在将来的工作中,将更加注重这方面的研究.

参考文献:

- [1] Rafael M M, Gabriel F A, Henry H, et al. Multidimensional projections for visual analysis of social networks. *Journal of Computer Science and Technology*, 2012, 4(4): 791–810.
- [2] Weslynn S A, Ariana C B. Assessing the “short mental distance” in eco-industrial networks. *Journal of Industrial Ecology*, 2012, 16(1): 70–82.
- [3] Raymond L P, Jennifer H G. Blind dates and arranged marriages: Longitudinal processes of network orchestration. *Organization Studies*, 2013, 34(11): 1623–1653.
- [4] 王宗尧, 隋 聪. 银行间网络模型与系统风险的分布式预警策略. *系统工程学报*, 2016, 31(6): 840–849.
Wang Z Y, Sui C. Interbank network model and distributed prediction strategy of systemic risk. *Journal of Systems Engineering*, 2016, 31(6): 840–849. (in Chinese)
- [5] Fidel A. Trees of the essential economic structures: A qualitative input-output method. *Journal of Regional Science*, 2006, 46(2): 333–353.
- [6] 刘 刚, 郭 敏. 中国宏观经济多部门网络及其性质的实证研究. *经济问题*, 2009, 18(2): 31–34.
Liu G, Guo M. An empirical research on Chinese macroeconomic multi-sector networks. *On Economic Problems*, 2009, 18(2): 31–34. (in Chinese)
- [7] 王茂军, 柴 箐. 北京市产业网络结构特征与调节效应. *地理研究*, 2013, 32(3): 543–555.
Wang M J, Chai Q. Beijing industrial network structural features and moderating effect. *Geographical Research*, 2013, 32(3): 543–555. (in Chinese)
- [8] 安琪儿, 安海忠, 王 朗. 中国产业间隐含能源流动网络分析. *系统工程学报*, 2014, 29(6): 754–762.
An Q E, An H Z, Wang L. Analysis of the embodied energy flow network between Chinese industries. *Journal of Systems Engineering*, 2014, 29(6): 754–762. (in Chinese)
- [9] 刘向荣, 杨建梅, 谢伟聪, 等. 价格波动关联效应投射网络拓扑性质研究: 以广东省为例. *系统工程学报*, 2014, 29(4): 433–442.
Liu X R, Yang J M, Xie W C, et al. The topological properties study of projected network on price volatility associated effect: Taking Guangdong as a case. *Journal of Systems Engineering*, 2014, 29(4): 433–442. (in Chinese)
- [10] 陈 曦, 席强敏, 李国平. 制造业内部产业关联与空间分布关系的实证研究. *地理研究*, 2015, 34(10): 1943–1956.
Chen X, Xi Q M, Li G P. Industrial linkage and spatial distribution of manufacturing industry. *Geographical Research*, 2015, 34(10): 1943–1956. (in Chinese)
- [11] 吴晓波, 姜雁斌. 经济转型: 基于网络分析的产业部门角色演化. *科学学研究*, 2010, 28(2): 243–249.
Wu X B, Jiang Y B. Economic transition mode identification based on network analysis. *Studies in Science of Science*, 2010, 28(2): 243–249. (in Chinese)
- [12] 王铜安. 基于社会网络视角的产业结构总体特征研究. *科研管理*, 2014, 35(7): 124–129.
Wang T A. Research on the general feature of Chinese industrial structure based on social network analysis. *Science Research Management*, 2014, 35(7): 124–129. (in Chinese)
- [13] 杨智峰, 陈霜华, 汪 伟. 中国产业结构变化的动因分析: 基于投入产出模型的实证研究. *财经研究*, 2014, 40(9): 38–49.
Yang Z F, Chen S H, Wang W. On motivations for industrial structure change: Empirical study based on input-output model. *Journal of Finance and Economics*, 2014, 40(9): 38–49. (in Chinese)
- [14] 杜华东, 赵尚梅. G7 与BRIC 产业结构演进的比较分析: 基于1995—2005年投入产出数据的网络分析. *管理工程学报*, 2014, 28(2): 17–25.
Du H D, Zhao S M. Empirical research on industrial structure evolution of G7 and BRIC: Network analysis of input-output data from 1995 to 2005. *Journal of Industrial Engineering and Engineering Management*, 2014, 28(2): 17–25. (in Chinese)
- [15] 温丹辉, 王 灿. 中国清洁能源的企业减排成本和社会减排成本. *系统工程学报*, 2014, 29(5): 592–601.
Wen D H, Wang C. Private and social mitigation cost comparisons of clean energies in China. *Journal of Systems Engineering*, 2014, 29(5): 592–601. (in Chinese)
- [16] Wasserman S, Faust K. *Social Network Analysis: Methods and Applications*. New York: Cambridge University Press, 1994.
- [17] Zhu Y X, Lu L, Zhang Q M, et al. Uncovering missing links with cold ends. *Physica A: Statistical Mechanics and Its Applications*, 2012, 391(22): 5769–5778.

- [18] Kaya M, Alhajj R. Development of multidimensional academic information networks with a novel data cube based modeling method. *Information Sciences*, 2014, 265(5): 211–224.
- [19] Buket K, Mustafa P. Age-series based link prediction in evolving disease networks. *Computers in Biology and Medicine*, 2015, 63(1): 1–10.
- [20] Yu H, Braun P, Yildirim M A, et al. High-quality binary protein interaction map of the yeast interactome network. *Science*, 2008, 322(1): 104–110.
- [21] Goldberg D S, Roth F P. Assessing experimentally derived interactions in a small world. *Proceeding of the National Academy of Science*, 2003, 100(8): 4372–4376.
- [22] Manning C D, Raghavan P, Schütze H. *Introduction to Information Retrieval*. New York: Cambridge University Press, 2008: 151–175.
- [23] Chen H, Li X, Huang Z. *Link Prediction Approach to Collaborative Filtering*. New York: ACM Press, 2005: 141–142.
- [24] Zhou S, Mondragón R J. Accurately modeling the internet topology. *Physical Review E*, 2004, 70(6): 1–20.
- [25] Juszczyszyn K, Gonczarek A, Tomczak J, et al. A Probabilistic Approach to Structural Change Prediction in Evolving Social Networks. New York: ACM Press, 2012: 996–1001.
- [26] Lu L Y, Jin C H, Zhou T. Similarity index based on local paths for link prediction of complex networks. *Physical Review E*, 2009, 80(2): 1–9.
- [27] Liu W P, Lu L Y. Link prediction based on local random walk. *Europhysics Letters*, 2010, 89(4): 1–6.
- [28] Murata T, Moriyasu S. Link Prediction of Social Networks Based on Weighted Proximity Measures. New York: ACM Press, 2007: 85–88.
- [29] Aaron C, Cristopher M, Newman M E J. Hierarchical structure and the prediction of missing links in networks. *Nature*, 2008, 453(5): 98–101.
- [30] Chungmok L, Minh P, Myong K J, et al. A network structural approach to the link prediction problem. *Inform Journal on Computing*, 2015, 27(2): 249–267.
- [31] Chen B, Chen L, Li B. A fast algorithm for predicting links to nodes of interest. *Information Sciences*, 2016, 329(1): 552–567.
- [32] Newman M E J. Clustering and preferential attachment in growing networks. *Physical Review E*, 2001, 64(2): 1–13.
- [33] Lu L Y, Zhou T. Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and Its Applications*, 2001, 390(6): 1150–1170.
- [34] Adamic L A, Adar E. Friends and neighbors on the web. *Social Networks*, 2003, 25(3): 211–230.
- [35] 吕琳媛. 复杂网络链路预测. *电子科技大学学报*, 2010, 39(5): 651–661.
Lü L Y. Link prediction on complex network. *Journal of University of Electronic Science and Technology of China*, 2010, 39(5): 651–661. (in Chinese)
- [36] 王秀英, 柴天佑, 郑秉霖. 炼钢-连铸调度模型参数优化设计方法. *系统工程学报*, 2011, 26(4): 531–537.
Wang X Y, Chai T Y, Zhen B L. Parameter optimization setting of steel making and continuous casting scheduling model. *Journal of Systems Engineering*, 2011, 26(4): 531–537. (in Chinese)
- [37] 夏 明, 张红霞. 投入产出分析: 理论、方法与数据. 北京: 中国人民大学出版社, 2013: 312–325.
Xia M, Zhang H X. *Input-output Analysis: Theory, Methods and Data*. Beijing: China Renmin University Press, 2013: 312–325. (in Chinese)
- [38] 许小可, 许 爽, 朱郁筱, 等. 复杂网络中链路的可预测性. *复杂系统与复杂性科学*, 2014, 11(1): 41–47.
Xu X K, Xu S, Zhu Y X, et al. Link predictability in complex network. *Complex Systems and Complexity Science*, 2014, 11(1): 41–47. (in Chinese)

作者简介:

王 斌(1974—), 男, 安徽桐城人, 博士, 讲师, 研究方向: 复杂网络, 系统建模, Email: wangpu2006@126.com;

王文平(1966—), 女, 山东日照人, 博士, 教授, 博士生导师, 研究方向: 复杂网络, 产业生态, Email: 101004452@seu.edu.cn;

费为银(1963—), 男, 安徽芜湖人, 博士, 教授, 博士生导师, 研究方向: 随机控制及应用, Email: wyfei@ahpu.edu.cn.