不确定属性图的子图同构及其判定算法

张春英 张 雪
(河北联合大学理学院 唐山 063009)

摘 要 在分析了复杂网络(社会网络)结构的基础上,针对不确定属性图的特征,首先定义了不确定属性图的期望子图同构;由于其只用一个阈值作为限制条件,虽然方法简单,但计算量大,故接着给出了不确定属性图的 α-β 子图同构的定义,并对其语义进行了解释说明;第三,设计并实现了子图同构算法;最后,通过实验证明 α-β 子图同构优于期望子图同构,同时分析了不同阈值情况下 α-β 子图同构的变化规律。α-β 子图同构算法的研究为不确定属性图的子图查询和社区挖掘工作奠定了基础。

关键词 不确定属性图,期望子图同构,α-β 子图同构
中图法分类号 TP311 **文献标识码** A

Uncertain Attribute Graph Sub-graph Isomorphism and its Determination Algorithm

ZHANG Chun-ying ZHANG Xue
(College of Science, Hebei United University, Tangshan 063009, China)

Abstract The uncertain attribute graph expectative sub-graph isomorphism is based on the analysis of complex network structure and the characteristic of uncertain attribute graph. The uncertain attribute graph expectative sub-graph isomorphism is only one threshold value as constraint conditions. The method is simple, but the computation is large amount. Therefore, it brought in the definition of α-β sub-graph isomorphic of uncertain attribute graph, explained the semantic, and designed and implemented the algorithm of α-β sub-graph isomorphism. Through the experiments was proved that α-β sub-graph isomorphic is better than expectative sub-graph, and it analyzed the variation in the different threshold cases. The research of α-β sub-graph isomorphism algorithm lays the foundation for uncertain attribute graph sub-graph query and community mining.

Keywords Uncertain attribute graph, Expectative sub-graph isomorphism, α-β sub-graph isomorphism

1 引言

作为一种通用的数据结构[1],图可以建模和表示现实世界中各种复杂的数据实体及实体之间的关系。在社会网络中,整个社会关系可以抽象为一个图,顶点表示社会中的个体,边表示个体之间的关系,社区表示为一个子图[3]。属性图[4]是在传统图的基础上考虑了结点和边的属性以及属性之间关系的图结构,可以更好地描述社会网络中各个结点及其之间关系的属性,更易于分析这些属性对网络结构的影响。然而,在社会网络或其他复杂网络中,结点与结点关系以及它们的属性都可能以一定的概率存在,不确定属性图则是在考虑此因素的情况下提出来的,它进一步描述了社会网络的不确定性[5]。

子图同构是对网络具有相同特性的子图的一种分类,研究图数据上的子图同构是对社会网络的一种更深刻的研究。目前,关于子图同构的研究工作有:董安国等给出了关于图模式挖掘中的子图同构算法,利用度序列和特征值构造了两种子图同构的算法来判断有向图和无向图的子图同构[6];解春欣提出了子图同构的验证算法 OES,采用逐边检验的方法寻找子图同构映射,以确定查询图是否为某个数据图的子图,通过调整边的顺序提高算法的执行效率[7];刘波提出基于关系模型的子图同构检测算法设计与实现,提出关系图分解索引(RG-DD)算法,提高了算法的效率[8];另有 Fred DePiero 提出使用分辨率来进行子图同构算法设计[9]等。目前这些子图同构算法均是基于传统图或不确定图设计的,而在实际的复杂网络中往往要用到不确定属性图来描述,如何针对不确定属性图设计不确定属性子图同构算法,是急需解决的重要问题之一。

本文在不确定属性图基础上,给出不确定属性图数据间子图同构的两种定义——期望子图同构和 α-β 子图同构。在概率条件下,不确定属性图的期望子图同构是确定图的子图同构定义的直接扩展,通过比较蕴含概率的期望值是否超过指定阈值进行同构判定。不确定属性图的期望子图同构的概率意义十分直观,但是存在计算代价巨大和匹配结果描述复杂的不足。不确定属性图 α-β 子图同构判定是利用两个限制阈值来替代不确定属性图期望子图同构判定中的单一限定阈值。

到稿日期:2012-09-21 返修日期:2012-12-30 本文受河北省自然科学基金(F2012209019)资助。
张春英 女,教授,硕士生导师,主要研究方向为为模糊集、粗糙集、计算机等,E-mail:48383107@qq.com;张 雪 女,硕士,主要研究方向为模糊集、粗糙集、人工智能。

本文第1节介绍不确定属性图;第2节是不确定属性图的期望子图的同构;第3节是不确定属性图的α-β子图同构;第4节是不确定属性图的α-β子图同构判定;第5节是实验分析;最后是结果与展望。

2 不确定属性图介绍

在属性图中,边和顶点都有各自的属性,由于信息的不确定性,边和顶点都具有不确定性,且属性也有不确定性,以此分为以下几方面来讨论,本文主要考虑简单无向图。

2.1 不确定属性图

定义1(不确定Ⅰ型属性图) 三元组$GA_{\mathrm{I}}=((V(VA,LV),E(EA,LE)),P_E,P_V)$,其中$(V(VA,LV),E(EA,LE))$是一个属性图,$P_E:E^{\to}[0,1]$是边的概率函数,其中$\forall e_i\in E$存在的概率用$P(e_i)$表示。$P_V:V^{\to}[0,1]$是顶点的概率函数,其中$\forall v_i\in V$存在的概率用$P(v_i)$表示,则称这个三元组为不确定Ⅰ型属性图。

不确定Ⅰ型属性图强调的是边和顶点的不确定性,且边和顶点概率函数是相互独立的,它们的属性是确定的。

性质1 当$P_E=1,P_V=1$时,不确定Ⅰ型属性图$GA_{\mathrm{I}}=((V(VA,LV),E(EA,LE)),P_E,P_V)$就退化成属性图$GA=(V(VA,LV),E(EA,LE))$。

定义2(不确定Ⅱ型属性图) 不确定Ⅱ型属性图定义为一个三元组$GA_{\mathrm{II}}=((V(VA,LV),E(EA,LE)),P_{EA},P_{VA})$,其中,$(V(VA,LV),E(EA,LE))$是一个属性图,$P_{EA}:EA^{\to}[0,1]$是边属性值的概率函数,其中$\forall ea_i\in EA$的值的概率用$P(ea_i)$表示,$P_{VA}:VA^{\to}[0,1]$是顶点属性值的概率函数,其中$\forall va_i\in VA$的值的概率用$P(ea_i)$表示。

不确定Ⅱ型属性图强调的是边和顶点的属性的不确定性,概率函数是相互独立的,认为它们的属性是确定的。

性质2 当$P_{EA}=1,P_{VA}(u,v)=1$时,不确定Ⅱ型属性图$GA_{\mathrm{II}}=((V(VA,LV),E(EA,LE)),P_{EA},P_{VA})$就退化成属性图$GA=(V(VA,LV),E(EA,LE))$。

定义3(不确定属性图) 不确定Ⅰ型属性图和不确定Ⅱ型属性图统称为不确定属性图,记为$GA_P=((V(VA,LV),E(EA,LE)),P)$,其中,$(V(VA,LV),E(EA,LE))$表示属性图,$P$是边、顶点及它们属性的概率函数($P$包括$P_E:E^{\to}[0,1]$,$P_V:V^{\to}[0,1]$,$P_{EA}:EA^{\to}[0,1]$,$P_{VA}:VA^{\to}[0,1]$)。

性质3 当$P=1$时,不确定属性图$GA_P=((V(VA,LV),E(EA,LE)),P)$就退化成属性图$GA=(V(VA,LV),E(EA,LE))$。

定义4 在不确定属性图$GA_P=((V(VA,LV),E(EA,LE)),P)$中,$VA=\{va_1,va_2,\cdots,va_{|VA|}\}$为$V$上的属性集合,除了顶点标识属性外,能唯一区分一个结点的属性称为顶点的关键属性。

定义5 在不确定属性图$GA_P=((V(VA,LV),E(EA,LE)),P)$中,在边集$E$上有属性集合$EA=\{ea_1,ea_2,\cdots,ea_{|EA|}\}$,除了顶点属性外,唯一能标识一条边的属性为边的关键属性。

结论1 不确定属性图是属性图的推广,属性图是不确定属性图的特例。

定义6(不确定属性子图) 不确定属性图$GA_P=((V(VA,LV),E(EA,LE)),P)$,其中属性图是$(V(VA,LV),E(EA,LE))$,$P$为概率函数;不确定属性图$GA_P'=((V'(VA,LV),E'(EA,LE)),P')$,其中属性图$(V'(VA,LV),E'(EA,LE))$,$P'$为包含顶点、边以及它们的属性的概率函数,如果$V(GA_P)\subseteq V(GA_P')$,$E(GA_P)\subseteq E(GA_P')$,$P_V\geqslant P_V'$,$P_{VA}\geqslant P_{VA}'$,$P_E\geqslant P_E'$,$P_{EA}\geqslant P_{EA}'$则称不确定属性图$GA_P$是不确定属性图$GA_P'$的属性子图,记为$GA_P\subseteq GA_P'$。

3 不确定属性图的期望子图同构

3.1 不确定Ⅰ属性图的期望子图同构

首先,我们来讨论不确定Ⅰ型属性图。

性质4 根据P值是否为1,可以把不确定Ⅰ型属性图划分成彼此不相交的子集$GA_{\mathrm{I}}^{UC}=\{GA_{\mathrm{I}}'\mid GA_{\mathrm{I}}'\subseteq GA_{\mathrm{I}},0<P(e_i)<1\text{ or }0<P(v_i)<1\text{ or }(0<P(e_i)<1\text{ and }0<P(v_i)<1),e_i\in E,v_i\in V\}$和$GA_{\mathrm{I}}^{C}=\{GA_{\mathrm{I}}''\mid GA_{\mathrm{I}}''\subseteq GA_{\mathrm{I}},P(e_i)=1,P(v_i)=1,e_i\in E,v_i\in V\}$,其中$GA_{\mathrm{I}}^{C}$是$GA_{\mathrm{I}}$的确定属性图集,$GA_{\mathrm{I}}^{UC}$是$GA_{\mathrm{I}}$的不确定属性图集,两个集合满足$GA_{\mathrm{I}}^{C}\cap GA_{\mathrm{I}}^{UC}=\Phi$,且$GA_{\mathrm{I}}^{C}\cup GA_{\mathrm{I}}^{UC}=GA_{\mathrm{I}}$。

不确定Ⅰ型属性图$GA_{\mathrm{I}}=((V(VA,LV),E(EA,LE)),P_E,P_V)$,它的一个可能世界确定属性图为$GA_{\mathrm{IC}}(V_1(VA,LV),E_1(EA,LE))$,其中$V_1(VA,LV)\subseteq V(VA,LV)$,$E_1(EA,LE)\subseteq E(EA,LE)$。$GA_{\mathrm{I}}$蕴含$GA_{\mathrm{IC}}$记为$GA_{\mathrm{I}}\Rightarrow GA_{\mathrm{IC}}$,则"$GA_{\mathrm{I}}$蕴含$GA_{\mathrm{IC}}$"的概率$P(GA_{\mathrm{I}}\Rightarrow GA_{\mathrm{IC}})$可依下述公式进行计算:

$$P(GA_{\mathrm{I}}\Rightarrow GA_{\mathrm{IC}})=\prod_{GA_{\mathrm{IC}}}P(e_i)\cdot P(v_i)\prod_{GA_{\mathrm{I}}^{UC}-GA_{\mathrm{IC}}}(1-P(e_i))(1-P(v_i)) \tag{1}$$

不确定Ⅰ型属性图$GA_{\mathrm{I1}},GA_{\mathrm{I2}}$,设$GA_{\mathrm{I1}}$在实际中以确定属性图$GA_1$的形式存在,$GA_{\mathrm{I2}}$在实际中以确定属性图$GA_2$的形式存在。"$GA_{\mathrm{I1}}$子图同构于$GA_{\mathrm{I2}}$"当且仅当$GA_1$子图同构于$GA_2$,因为$GA_1$和$GA_2$是未知的,$GA_1$是可能世界图集$s(GA_{\mathrm{I1}})$中任一确定属性图,$GA_2$是可能世界图集$s(GA_{\mathrm{I2}})$中任一确定属性图,所以$GA_{\mathrm{I1}}$是否子图同构于$GA_{\mathrm{I2}}$无法用简单的"是"或"否"回答。依据式(2)可计算"GA_{I1}子图同构于GA_{I2}"的概率$P(GA_{\mathrm{I1}}\subseteq GA_{\mathrm{I2}})$:

$$P(GA_{\mathrm{I1}}\subseteq GA_{\mathrm{I2}})=\sum_{GA_1\cap GA_2}P(GA_{\mathrm{I1}}\Rightarrow GA_1)*P(GA_{\mathrm{I2}}\Rightarrow GA_2)*\varphi(GA_1,GA_2) \tag{2}$$

函数$\varphi(GA_1,GA_2)$值域为$\{0,1\}$,如果GA_1与GA_2子图同构,则$\varphi(GA_1,GA_2)$值为1,否则为0,显然,$P(GA_{\mathrm{I1}}\subseteq GA_{\mathrm{I2}})$是$\varphi(GA_1,GA_2)$的概率期望值。根据此期望值可以定义不确定Ⅰ型属性图期望子图同构。

定义7 已知不确定Ⅰ型属性图GA_{I1}和GA_{I2},期望阈值$\delta\in(0,1]$,GA_{I1}期望子图同构于GA_{I2},当且仅当$P(GA_{\mathrm{I1}}\subseteq GA_{\mathrm{I2}})\geqslant\delta$,记为$GA_{\mathrm{I1}}\subseteq_{\delta}GA_{\mathrm{I2}}$。

3.2 不确定Ⅱ型属性图期望子图同构

同理,可以定义不确定Ⅱ型属性图的期望子图同构。

性质5 根据P值是否为1,可以把不确定Ⅱ型属性图划分成彼此不相交的子集$GA_{\mathrm{II}}^{UC}=\{GA_{\mathrm{II}}'\mid GA_{\mathrm{II}}'\subseteq GA_{\mathrm{II}},0<P(ea_i)<1\text{ or }0<P(va_i)<1\text{ or }(0<P(ea_i)<1\text{ and }0<P(va_i)<1),ea_i\in EA,va_i\in VA\}$和$GA_{\mathrm{II}}^{C}=\{GA_{\mathrm{II}}''\mid GA_{\mathrm{II}}''\subseteq GA_{\mathrm{II}},P(ea_i)=1,P(va_i)=1,ea_i\in EA,va_i\in VA\}$,其中$GA_{\mathrm{II}}^{C}$是$GA_{\mathrm{II}}$的确定属性图集,$GA_{\mathrm{II}}^{UC}$是$GA_{\mathrm{II}}$的不确定属性图集,两个集合满足$GA_{\mathrm{II}}^{C}\cap GA_{\mathrm{II}}^{UC}=\Phi$,且$GA_{\mathrm{II}}^{C}\cup GA_{\mathrm{II}}^{UC}=GA_{\mathrm{II}}$。

已知不确定Ⅱ型属性图$GA_{\mathrm{II}}=((V(VA,LV),E(EA,

$LE)), P_{EA}, P_{VA})$，它的一个可能世界属性图为 $GA_{\text{II}C}(V_1(VA, LV), E_1(EA, LE))$，其中 $V_1(VA, LV) \subseteq V(VA, LV)$，$E_1(EA, LE) \subseteq E(EA, LE)$。$GA_{\text{II}}$ 蕴含 $GA_{\text{II}C}$，记为 $GA_{\text{II}} \Rightarrow GA_{\text{II}C}$，则“$GA_{\text{II}}$ 蕴含 $GA_{\text{II}C}$”的概率 $P(GA_{\text{II}} \Rightarrow GA_{\text{II}C})$ 可依下述公式进行计算：

$$P(GA_{\text{II}} \Rightarrow GA_{\text{II}C}) = \prod_{GA_{\text{II}C}} P(ea_i) \cdot P(va_i) \prod_{GA_{\text{II}}^{UC} - GA_{\text{II}C}} (1 - P(ea_i)) \cdot (1 - P(va_i)) \quad (3)$$

已知不确定 II 属性图 GA_{III}，GA_{II2}，设 GA_{III} 在实际中以确定属性图 GA_1 的形式存在，GA_{II2} 在实际中以确定属性图 GA_2 的形式存在。“GA_{III} 子图同构于 GA_{II2}”当且仅当 GA_1 子图同构于 GA_2，因为 GA_1 和 GA_2 是未知的，GA_1 是可能世界图集 $s(GA_{\text{III}})$ 中任一确定属性图，GA_2 是可能世界图集 $s(GA_{\text{II2}})$ 中任一确定属性图，所以 GA_{III} 是否子图同构于 GA_{II2} 也无法用简单的“是”或“否”回答，依据式(4)可计算“GA_{III} 子图同构于 GA_{II2}”的概率 $P(GA_{\text{III}} \subseteq GA_{\text{II2}})$：

$$P(GA_{\text{III}} \subseteq GA_{\text{II2}}) = \sum_{GA_1 \cap GA_2} P(GA_{\text{III}} \Rightarrow GA_1) * P(GA_{\text{II2}} \Rightarrow GA_2) * \varphi(GA_1, GA_2) \quad (4)$$

函数 $\varphi(GA_1, GA_2)$ 值域为 $\{0,1\}$，如果 GA_1 与 GA_2 子图同构，则 $\varphi(GA_1, GA_2)$ 值为 1，否则为 0，显然，$P(GA_{\text{III}} \subseteq GA_{\text{II2}})$ 是 $\varphi(GA_1, GA_2)$ 的概率期望值。根据此期望值可以定义不确定 II 型属性图期望子图同构。

定义 8 已知不确定 II 型属性图 GA_{III}，GA_{II2}，期望阈值 $\delta \in (0,1]$，GA_{III} 期望子图同构于 GA_{II2}，当且仅当 $P(GA_{\text{III}} \subseteq GA_{\text{II2}}) \geqslant \delta$，记为 $GA_{\text{II2}} \subseteq_{\delta} GA_{\text{II2}}$。

3.3 不确定属性图的期望子图同构

定义 9(不确定属性图的期望子图同构) 已知不确定 I 型属性图 GA_{P1}，GA_{P2}，期望阈值 $\delta \in (0,1]$，GA_{P1} 期望子图同构于 GA_{P2}，当且仅当 $\sqrt{P(GA_{\text{III}} \subseteq GA_{\text{II2}}) P(GA_{\text{II}} \subseteq GA_{\text{I2}})} \geqslant \delta$，记为 $GA_{P2} \subseteq_{\delta} GA_{P2}$。

判定算法计算代价巨大，在应用时存在匹配结果描述困难的问题，这些问题都限制了期望子图同构的应用。

4 不确定属性图的 α-β 子图同构

4.1 不确定 I 属性图的 α-β 子图同构

定义 10 不确定 I 型属性图 GA_{I} 的不确定属性集 $GA_{\text{I}}^{UC} = \{GA \mid GA \subseteq GA_{\text{I}}, 0 < P_v < 1, 0 < P_e < 1\}$，函数 $\alpha: 2^{GA_{\text{I}}^{UC}} \to (0,1]$ 称为误差函数，对于 GA_{I}^{UC} 的任一子图 GA_{I}'，误差函数值 $\alpha(GA_{\text{I}\Delta})$ 为：

$$\alpha(GA_{\text{I}}') = \begin{cases} 1 - \prod_{GA_{\text{I}}' \subseteq GA_{\text{I}}} (1 - P(e_i))(1 - P(v_i)), & GA_{\text{I}}' \neq \Phi \\ 0, & GA_{\text{I}}' = \Phi \end{cases} \quad (5)$$

定义 11 不确定 I 型属性图 GA_{I} 的不确定属性集为 GA_{I}^{UC}，其中 $GA_{\text{I}\Delta}$ 是 GA_{I} 的任一子集，函数 $\beta: 2^{GA_{\text{I}\Delta}} \to (0,1]$ 称为强度函数，则

$$\beta(GA_{\text{I}\Delta}) = \sqrt[|GA_{\text{I}\Delta}|]{\prod_{GA_{\text{I}\Delta} \in GA_{\text{I}}} p(v_i) \cdot P(e_i)} \quad (6)$$

为了表述的简便，我们做如下规定，$I_{\max}(GA_{\text{I}})$ 是不确定 I 型属性图 GA_{I} 蕴含的最大确定属性图，$I_{\max}(GA_{\text{I}}) - GA_{\text{I}}'$ 表示在确定属性图 $I_{\max}(GA_{\text{I}})$ 中删去所有 GA_{I}' 的确定属性图。函数 $f(E_1)$ 表示在不确定 I 型属性图 $GA_{\text{II}} = ((V_1(VA, LV), E_1(EA, LE)), P_{E_1}, P_{V_1})$ 中 E_1 所对应的边的集合，即 $f(E_1) = \{f(e) \mid e \in E_1\}$，函数 $f(V_1)$ 表示在不确定 I 型属性图 $GA_{\text{II}} = ((V_1(VA, LV), E_1(EA, LE)), P_{E_1}, P_{V_1})$ 中 V_1 所对应的顶点的集合，即 $f(V_1) = \{f(v) \mid v \in V_1\}$。

定义 12 已知不确定 I 型属性图 GA_{II} 和 GA_{I2}，设 GA_{II} 的不确定属性集为 GA_{I}^{UC}，如果存在 $GA_{\text{I}}' \subseteq GA_{\text{I}}^{UC}$ 和映射函数 f 同时满足：

(1) $cI_{\max}(GA_{\text{II}}) - GA_{\text{I}}'$ 子图同构于 $I_{\max}(GA_{\text{I2}})$，f 为同构映射函数；

(2) $\alpha(GA_{\text{I}}') \leqslant \alpha_{\max}$，常数 $\alpha_{\max}$ 称为误差阈值，$\alpha_{\max} \in [0,1]$；

(3) $\beta(f(E_1) f(V_1)) \geqslant \beta_{\min}$，常数 $\beta_{\min}$ 称为强度阈值，$\beta \in [0,1]$。

那么不确定 I 型属性图 GA_{II}“α-β 子图同构于”不确定 I 型属性图 GA_{I2}，记为 $GA_{\text{II}} \subseteq_{\alpha,\beta} GA_{\text{I2}}$。

4.2 不确定 I 型属性图的 α-β 子图同构的语义

条件 1 “$I_{\max}(GA_{\text{I}}) - GA_{\text{I}}'$ 子图同构于 $I_{\max}(GA_{\text{I2}})$”可表示为样本空间 $s(GA_{\text{II}}, GA_{\text{I}}')$ 中每个图都是确定图 $I_{\max}(GA_{\text{I2}})$ 的子图，集合 $s(GA_{\text{II}})$ 是被 GA_{II} 蕴含的所有确定图集合，GA_{I}' 将 $s(GA_{\text{II}})$ 分为 $s(GA_{\text{II}}, GA_{\text{I}}')$ 和 $s(GA_{\text{II}}) \backslash s(GA_{\text{II}}, GA_{\text{I}}')$，其中 $s(GA_{\text{II}}, GA_{\text{I}}') = \{I \mid I \subseteq I_{\max}(GA_{\text{II}}) - GA_{\text{I}}'\}$。

条件 2 “$\alpha(GA_{\text{I}}') \leqslant \alpha_{\max}$”对 GA_{I}' 进行限制，为了保证“不确定 I 型属性图 GA_{II} 蕴含子样本空间 $s(GA_{\text{II}}, GA_{\text{I}}')$ 的概率足够大，即 $P(GA_{\text{I}} \Rightarrow s(GA_{\text{II}}, GA_{\text{I}}'))$ 的概率足够大，根据式(1)计算得 $P(GA_{\text{I}} \Rightarrow s(GA_{\text{II}}, GA_{\text{I}}')) = \sum_{I \in s(GA_{\text{I}}, GA_{\text{I}}')} P(GA_{\text{I}} \Rightarrow I) \geqslant 1 - \alpha(GA_{\text{I}}')$。

条件 3 “$\beta(f(E_1) f(V_1)) \geqslant \beta_{\min}$”对同构函数 f 进行限制，以保证不确定 I 型属性图 GA_{I2} 蕴含 $I_{\max}(GA_{\text{II}}) - GA_{\text{I}}'$ 足够大。所以 $I_{\max}(GA_{\text{I2}})$ 的子图 $I_{\max}'(GA_{\text{I2}})$ 和 $I_{\max}(GA_{\text{II}}) - GA_{\text{I}}'$ 同构，其中 $I_{\max}'(GA_{\text{I2}})$ 是以 $f(E_1)$ 为边、以 $f(V_1)$ 为顶点的子图。

定理 1 如果不确定 I 型属性图 GA_{II} 在 $\alpha_{\max}$ 和 $\beta_{\min}$ 的限制下同构于不确定 I 型属性图 GA_{I2}，那么 $P(GA_{\text{II}} \subseteq GA_{\text{I2}}) \geqslant \delta(\alpha_{\max}, \beta_{\min})$，其中 $\delta(\alpha_{\max}, \beta_{\min}) = (1 - \alpha_{\max}) * \beta_{\min}$。

证明：不确定 I 型属性图 GA_{II} 和 GA_{I2} α-β 子图同构，则满足 $\alpha(GA_{\text{I}}') \leqslant \alpha_{\max}$ 和 $\beta(f(E_1) f(V_1)) \geqslant \beta_{\min}$，且 $P(GA_{\text{II}} \subseteq GA_{\text{I2}}) = \sum_{GA_1 \cap GA_2} P(GA_{\text{II}} \Rightarrow GA_1) * P(GA_{\text{I2}} \Rightarrow GA_2) * \varphi(GA_1, GA_2)$，其中，$P(GA_{\text{I}} \Rightarrow GA_{\text{I}C}) = \prod_{GA_{\text{I}C}} P(e_i) \cdot P(v_i) \prod_{GA_{\text{I}}^{UC} - GA_{\text{I}C}} (1 - P(e_i))(1 - P(v_i))$，由于误差函数 $\alpha(GA_{\text{I}}') = \begin{cases} 1 - \prod_{GA_{\text{I}}' \subseteq GA_{\text{I}}} (1 - P(e_i))(1 - P(v_i)), & GA_{\text{I}}' \neq \Phi \\ 0, & GA_{\text{I}}' = \Phi \end{cases}$，即 $1 - \alpha_{max}$ 由公式得 $\begin{cases} \prod_{GA_{\text{I}}' \subseteq GA_{\text{I}}} (1 - P(e_i))(1 - P(v_i)), & GA_{\text{I}}' \neq \Phi \\ 1, & GA_{\text{I}}' = \Phi \end{cases}$，强度函数为 $\beta(GA_{\text{I}\Delta}) = \sqrt[|GA_{\text{I}\Delta}|]{\prod_{GA_{\text{I}\Delta} \in GA_{\text{I}}} p(v_i) \cdot P(e_i)}$，故

$$(1 - \alpha_{\max}) * \beta_{\min} = \begin{cases} \sqrt[|GA_{\text{I}\Delta}|]{\prod_{GA_{\text{I}\Delta} \in GA_{\text{I}}} P(v_i) \cdot P(e_i)} * \\ \prod_{GA_{\text{I}}' \subseteq GA_{\text{I}}} (1 - P(e_i))(1 - P(v_i)) \\ \sqrt[|GA_{\text{I}\Delta}|]{\prod_{GA_{\text{I}\Delta}} P(v_i) P(e_i)} \end{cases}$$

所以 $P(GA_{\text{II}} \subseteq GA_{\text{I2}}) \geqslant \delta(\alpha_{\max}, \beta_{\min})$。

如果不确定 I 型属性图 GA_{II} α-β 子图同构于不确定 I 型属性图 GA_{I2}，则在现实世界中不确定 I 型属性图 GA_{II} 子图同

构于不确定 I 型属性图 GA_{I2} 的概率 $P(GA_{II}\subseteq GA_{I2})\geqslant\delta(\alpha_{max},\beta_{min})$，如定义 1 所示，期望子图同构使用单一常数 δ 作为限定阈值，α-β 子图同构使用两个常数 α_{max} 和 β_{min} 作为代替常数 δ 的限定阈值。α-β 子图同构具有概率意义，可表示为：如果不确定 I 型属性图 GA_{II} α-β 子图同构于不确定 I 型属性图 GA_{I2} 且 $\delta(\alpha_{max},\beta_{min})$ 不小于期望阈值 δ，那么不确定 I 型属性图 GA_{II} 期望子图同构于不确定 I 型属性图 GA_{I2}。

4.3 不确定 II 型属性图 α-β 子图同构

定义 13 不确定 II 型属性图 GA_{II} 的不确定属性图集为 $GA_{II}^{UC}=\{GA\mid GA\subseteq GA_{II},0<P_{VA}<1,0<P_{EA}<1\}$，函数 α：$2^{GA_{II}^{UC}}\rightarrow(0,1]$ 称为误差函数，对于 GA_{II}^{UC} 的任一子集 $GA_{II\Delta}$，误差函数值 $\alpha(GA_{II\Delta})$ 为：

$$\alpha((GA_{II\Delta})=\begin{cases}1-\prod_{GA_{II\Delta}\subseteq GA_{II}}(1-P(ea_i))(1-P(va_i)), & GA_{II\Delta}\neq\Phi\\ 0, & GA_{II\Delta}=\Phi\end{cases}\tag{7}$$

定义 14 不确定 II 型属性图 GA_{II} 的不确定属性图集为 GA_{II}^{UC}，其中 $GA_{II\Delta}$ 是 GA_{II} 的任一子集，函数称为强度函数，则

$$\beta(GA_{II\Delta})=\sqrt[|GA_{II\Delta}|]{\prod_{GA_{II\Delta}\in GA_{III UC}}p(va_i)\cdot P(ea_i)}\tag{8}$$

同不确定 I 型属性图类似，为了表述简便，做如下规定，$I_{max}(GA_{II})$ 是不确定 II 型属性图 GA_{II} 蕴含的最大确定属性图，$I_{max}(GA_{II})-GA_{II\Delta}$ 表示在确定属性图 $I_{max}(GA_{II})$ 中删去所有的 $GA_{II\Delta}$ 的确定属性图。函数 $f(EA_1)$ 表示在不确定 II 型属性图 $GA_{II}=((V_1(VA_1,LV),E_1(EA_1,LE)),P_{EA},P_{VA})$ 中 EA_1 所对应的边的属性的集合，即 $f(EA_1)=\{f(ea)\mid ea\in EA_1\}$，函数 $f(VA_1)$ 表示 VA_1 在不确定 II 型属性图 $GA_{II}=((V_1(VA_1,LV),E_1(EA_1,LE)),P_{EA},P_{VA})$ 中 VA_1 所对应的顶点的属性的集合，即 $f(VA_1)=\{f(va)\mid va\in VA_1\}$。

定义 15 已知不确定 II 型属性图 GA_{III} 和 GA_{II2}，设 GA_{III} 的不确定属性集为 GA_{II}^{UC}，如果存在 $GA_{II}'\subseteq GA_{II}^{UC}$ 和映射函数 f 同时满足：

(1) $I_{max}(GA_{III})-GA_{II}'$ 子图同构于 $I_{max}(GA_{II2})$，f 为同构映射函数；

(2) $\alpha(GA'_{II})\leqslant\alpha_{max}$，常数 α_{max} 称为误差阈值，$\alpha_{max}\in[0,1]$；

(3) $\beta(f(EA_1)f(VA_1))\geqslant\beta_{min}$，常数 β_{min} 称为强度阈值，$\beta\in[0,1]$。

那么不确定 II 型属性图 GA_{III} “α-β 子图同构于”不确定 II 型属性图 GA_{II2}，记为 $GA_{III}\subseteq_{\alpha,\beta}GA_{II2}$。

4.4 不确定 II 属性图的 α-β 子图同构的语义

条件 1 “$I_{max}(GA_{III})-GA_{II}'$ 子图同构于 $I_{max}(GA_{II2})$”可表示为样本空间 $s(GA_{III},GA_{II})'$ 中每个图都是确定图 $I_{max}(GA_{II2})$ 的子图，集合 $s(GA_{III})$ 是被 GA_{III} 蕴含的所有确定图集合，GA_{II}' 将 $s(GA_{III})$ 分为 $s(GA_{III},GA_{II}')$ 和 $s(GA_{III})\backslash s(GA_{III},GA_{II})'$，其中 $s(GA_{III},GA_{II}')=\{I\mid I\subseteq I_{max}(GA_{III})-GA_{II}'\}$。

条件 2 “$\alpha(GA_{II}')\leqslant\alpha_{max}$”对 GA_{II}' 进行限制，为了保证“不确定 II 型属性图 GA_{III} 蕴含子样本空间 $s(GA_{III},GA_{II}')$ 的概率足够大，即 $P(GA_{III}\Rightarrow s(GA_{III},GA_I'))$ 的概率足够大，根据式(1)计算得

$$P(GA_{III}\Rightarrow s(GA_{III},GA_I'))=\sum_{I\in s(GA_{II},GA_{II}')}P(GA_{II}\Rightarrow I)\geqslant 1-\alpha(GA_{II}')$$

条件 3 “$\beta(f(E_1)f(V_1))\geqslant\beta_{min}$”对同构函数 f 进行限制，以保证不确定 II 型属性图 GA_{II2} 蕴含 $I_{max}(GA_{III})-GA_{II}'$ 足够大。所以 $I_{max}(GA_{II2})$ 以 $f(EA_1)$ 为边以 $f(VA_1)$ 为顶点的属性的子图 $I'(f)$ 和 $I_{max}(GA_{III})-GA_{II}'$ 同构。

定理 2 如果不确定 II 型属性图 GA_{III} 在 α_{max} 和 β_{min} 的限制下同构于不确定 II 型属性图 GA_{II2}，那么 $P(GA_{III}\subseteq GA_{II2})\geqslant\delta(\alpha_{max},\beta_{min})$，其中 $\delta(\alpha_{max},\beta_{min})=(1-\alpha_{max})*\beta_{min}$

证明：不确定 II 型属性图 GA_{III} 和 GA_{II2} $\alpha-\beta$ 子图同构则满足 $\alpha(GA_{II}')\leqslant\alpha_{max}$，和 $\beta(f(EA_1)f(VA_1))\geqslant\beta_{min}$，且 $P(GA_{III}\subseteq GA_{II2})=\sum_{GA_1\cap GA_2}P(GA_{III}\Rightarrow GA_1)*P(GA_{II2}\Rightarrow GA_2)*\varphi(GA_1,GA_2)$，其中 $P(GA_{II}\Rightarrow GA_{IIC})=\prod_{GA_{IIC}}P(ea_i)\cdot P(va_i)\prod_{GA_{II}^{UC}-GA_{IIC}}(1-P(ea_i))(1-P(va_i))$，由于误差函数 $\alpha(GA_{II}')=\begin{cases}1-\prod_{GA_{II}'\subseteq GA_{II}}(1-P(ea_i))(1-P(va_i)), & GA_{II}'\neq\Phi\\ 0, & GA_{II}'=\Phi\end{cases}$，即 $1-\alpha_{max}$ 由公式得

$$\begin{cases}\prod_{GA_{II}\subseteq' GA_{II}}(1-P(ea_i))(1-P(va_i)), & GA_{II}'\neq\Phi\\ 1, & GA_{II}'=\Phi\end{cases}$$

强度函数为 $\beta(GA_{II\Delta})=\sqrt[|GA_{II\Delta}|]{\prod_{GA_{II\Delta}\in GA_{II}}p(va_i)\cdot P(ea_i)}$，因此

$$(1-\alpha_{max})*\beta_{min}=\begin{cases}\sqrt[|GA_{II\Delta}|]{\prod_{GA_{II\Delta}\in GA_{II}}P(va_i)\cdot P(ea_i)}\\ *\prod_{GA_{II}'\subseteq GA_{II}}(1-P(ea_i))(1-P(va_i)),\\ \sqrt[|GA_{II\Delta}|]{\prod_{GA_{II\Delta}}P(va_i)P(ea_i)}\end{cases}$$

所以 $P(GA_{III}\subseteq GA_{II2})\geqslant\delta(\alpha_{max},\beta_{min})$。

如果不确定 II 型属性图 GA_{III} α-β 子图同构于不确定 II 型属性图 GA_{II2}，则在现实世界中不确定 II 型属性图 GA_{III} 子图同构于不确定 II 型属性图 GA_{II2} 的概率 $P(GA_{III}\subseteq GA_{II2})\geqslant\delta(\alpha_{max},\beta_{min})$，如定义 2 所示，期望子图同构使用单一常数 δ 作为限定阈值，α-β 子图同构使用两个常数 α_{max} 和 β_{min} 作为代替常数 δ 的限定阈值。α-β 子图同构具有概率意义，可表示为：如果不确定 I 型属性图 GA_{II} α-β 子图同构于不确定 I 型属性图 GA_{II2} 且 $\delta(\alpha_{max},\beta_{min})$ 不小于期望阈值 δ，那么，不确定 II 型属性图 GA_{III} 期望子图同构于不确定 II 型属性图 GA_{II2}。

4.5 不确定属性图的 α-β 子图同构

定义 16(不确定属性图的 α-β 子图同构) 已知不确定属性图 GA_{P1} 和 GA_{P2}，设 GA_{P1} 的不确定属性集为 GA_{P1}^{UC}，如果存在 $GA_P'\subseteq GA_P^{UC}$ 和映射函数 f 同时满足：

(1) $I_{max}(GA_{P1})-GA_P'$ 子图同构于 $I_{max}(GA_{P2})$，f 为同构映射函数；

(2) $\sqrt{\alpha(GA_{II}')\alpha(GA_I')}\leqslant\alpha_{max}$，常数 α_{max} 称为误差阈值，$\alpha_{max}\in[0,1]$；

(3) $\sqrt{\beta(f(EA_1)f(VA_1)f(E_1)f(V_1))}\geqslant\beta_{min}$，常数 β_{min} 称为强度阈值，$\beta\in[0,1]$。

那么不确定属性图 GA_{P1} “α-β 子图同构于”不确定属性图 GA_{P2}，记为 $GA_{P1}\subseteq_{\alpha,\beta}GA_{P2}$

5 不确定属性图的 α-β 子图同构判定算法研究

本文以确定图上常用子图同构判定算法 ullmann[11] 为基础，设计实现不确定属性图 α-β 子图同构判定算法如下：

Adbumlgo(GA_{P1}, GA_{P2}, H, F, $\alpha(GA_I')$, $\alpha(GA'_{II})$, $\beta(GA_I')$, $\beta(GA'_{II})$)

Input: Uncertain attribute graph GA_{P1} with N_1 nodes

Uncertain attribute graph GA_{P2} with N_2 nodes
ArrayH[N_1],H[i]initializedby 0,$1\leqslant i\leqslant N_1$
ArrayF[N_2],F[i] initializedby 0,$1\leqslant i\leqslant N_2$
Integerd,initialized by 1,
$\alpha(GA_I')$,initialized by 0
$\alpha(GA_{II}')$,initialized by 0
$\beta(GA_I')$,initialized by 0
$\beta(GA_{II}')$,initialized by 0
Output:boolean:$GA_{P1}\subseteq_{\alpha,\beta}GA_{P1}$
(1) if $d>N_1$ then
(2) if $\beta(GA_I')<\beta_{min}$,or $\beta(GA_{II'})<\beta_{min}$then return false
(3) else return true
(4) end if
(5) end if
(6) for each $1\leqslant k\leqslant N_2 \wedge F[k]=0$ do
(7) if $I(d)\neq l(k)$then goto line6
(8) calculate and update $\alpha(GA_I')$,$\alpha(GA_{II}')$
(9) if $\alpha(GA_I')>\alpha_{max}$,or$\alpha(GA_{II}')$then
(10) recover $\alpha(GA_I')$,$\alpha(GA_{II}')$and goto line6
(11) end if
(12) H[d]=k
(13) F[k]=1
(14) calculate and update $\beta(GA_{I\Delta})$,$\beta(GA_{II\Delta})$
(15) ifAdvUmAlgo(GA_{Pq},GA_{P2},H,F,d+1) then
(16) return true
(17) end if
(18) F[k]=0
(19) recover $\alpha(GA_I')$,$\alpha(GA_{II})$,'$\beta(GA_I')$ and $\beta(GA_{II}')$
(20) end for
(21) return false

算法:AdvUmAlgo 是以深度递推搜索的方法来实现的,在深度为 d 的递归调用开始前,GA_{P1} 的前 $d-1$ 个顶点、边及属性已经匹配完毕,对应关系存储在数组 H 的前 $d-1$ 位中。在不确定属性图上的"$\alpha\beta$"子图同构匹配的精确算法中,如果当前数据图中没有合适顶点和边与查询图中当前节点相匹配,则尝试忽略查询图中当前节点和边的属性的匹配。

6 实验分析

6.1 实验数据和参数设置

不确定属性图数据包括查询图和不确定属性图,实验用 4 组查询图,GA5,GA8,GA11 和 GA14 是社会网络中个体之间的关系,其中每个图的顶点数为 n,不确定属性图数据之间的 $\alpha\beta$ 子图同构用到的参数主要有(α_{max},β_{max}),设定 3 组参数,分别是(0.2,0.7),(0.4,0.7),(0.4,0.3)。主要观察在强度阈值相同的情况下,随着误差阈值和计算的时间如何变化;在误差阈值相同的情况下,随着强度阈值和计算的时间如何变化。

6.2 实验数据和参数设置

通过实验,从分析图 1 可以看见在强度阈值相同的情况下,误差阈值越小,计算的时间越短;在误差阈值相同的情况下,随着强度阈值的增大,计算执行的时间增长。图 2 为优化前后的 $\alpha\beta$ 子图同构判定算法的平均判定执行时间,该项实验将参数固定为具有代表性的第二组数值,通过图 2 可知,判定算法在搜索顺序进行优化以后执行时间大幅缩短。

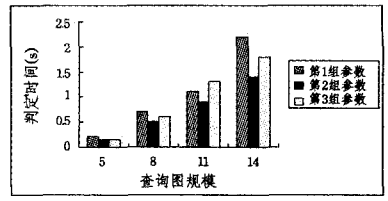

图 1

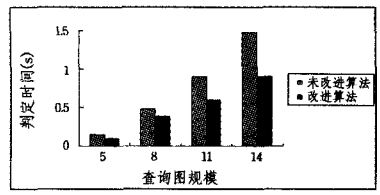

图 2

结束语 不确定属性图的期望子图同构是一种理想状态,但在实际问题中往往很难达到,文中重点阐述的是不确定属性图的 $\alpha\beta$ 子图的定义及其判定方法,在设定两个阈值的情况下,$\alpha\beta$ 子图同构更接近于实际,其判定算法的效率在阈值可调控的前提下也得到了大大提高。$\alpha\beta$ 子图同构的研究为复杂网络(社会网络)中子图查询和社区挖掘的工作深入开展奠定了基础。但是,如何设定阈值才能使 $\alpha\beta$ 子图同构更加合理,使其与实际情况更匹配,是进一步研究的问题。

参 考 文 献

[1] 李先通. 图数据查询技术的研究[D]. 哈尔滨:哈尔滨工业大学,2009

[2] 张硕,高宏,李建中. 不确定图数据库中高效查询处理[J]. 计算机学报,2009,32(10):2066-2079

[3] 张一楠,邹兆年,李建中. 不确定图间 $\alpha\beta$ 子图同构匹配算法[J]. 智能计算机与应用,2011(10):1-3

[4] Zhang Chun-ying, Guo Jing-feng, Chen Xiao. Research on random walk rough matching algorithm of attribute subgraph[C]// 2011 International Conference on Advanced Materials and Computer Science, ICAMCS 2011. Chengdu, China(EI), May 2011: 297-302

[5] Cheng Jian-kun, Huang Tong-shan. A subgraph isomorphism algorithm using resolution original research article[J]. Pattern Recognition, 1981, 13(5): 371-379

[6] 董安国,高琳,赵建邦. 图模式挖掘中的子图同构算法[J]. 数学的实践与认识,2011(7):105-111

[7] 解春欣,汪为. 子图同构验证算法OES[J]. 计算机工程,2011,2(3):30-32

[8] 刘波,房斌,张世勇,等. 基于关系模式的子图同构检测算法设计与实现[J]. 计算机工程,2011,6(11):62-63,66

[9] DePiero F, Krout D. An algorithm using length-r paths to approximate subgraph isomorphism[J]. Pattern Recognition Letters, 2003, 24(1-3): 33-46

[10] 周傲英,金澈清,王国仁,等. 不确定性数据管理技术研究综述[J]. 计算机学报,2009,32(1):1-16

[11] Bodic P L, Héroux P, Adam S, et al. An integer linear program for substitution-tolerant subgraph isomorphism and its use for symbol spotting in technical drawings Original Research Article [J]. Pattern Recognition, 2012, 45(12): 4214-4224