面向不确定图的概率可达查询

袁 野[1),2)] 王国仁[1),2)]

[1)] (医学影像计算教育部重点实验室(东北大学) 沈阳 110004)

[2)] (东北大学信息科学与工程学院 沈阳 110004)

摘 要 图的可达性查询被广泛应用于生物网络、社会网络、本体网络、RDF 数据库和 XML 数据库等. 由于对数据操作时引入的噪声和错误使这些图数据具有不确定性, 已经有大量的针对不确定 RDF 和 XML 数据库的研究. 文中使用可能世界语义模型构建不确定图, 基于该模型, 研究了概率可达查询(PR). 处理 PR 查询是# P 完全问题, 对此文中首先给出一个基本随机算法, 可快速地估算出可达概率, 并且该值有很高的精确度. 进一步, 文中为随机算法引入条件分布(称为"条件随机算法"), 采用图的不相交路径集和割集作为条件概率分布, 因此改进的随机算法可准确地并且是在多项式时间内处理查询. 最后基于真实不确定图数据的大量实验结果验证了文中的设计.

关键词 不确定图; 可能世界; 条件随机算法; 路径集; 割集

中图法分类号 TP311 **DOI** 号: 10.3724/SP.J.1016.2010.01378

Answering Probabilistic Reachability Queries over Uncertain Graphs

YUAN Ye[1),2)] WANG Guo-Ren[1),2)]

[1)] (*Key Laboratory of Medical Image Computing (Northeastern University), Ministry of Education, Shenyang* 110004)

[2)] (*College of Information Science and Engineering, Northeastern University, Shenyang* 110004)

Abstract Graph reachability queries are widely used in biological networks, social networks, ontology networks, RDF and XML databases. Meanwhile, data extracted from those applications is inherently uncertain due to noise, incompleteness and inaccuracy, and many works have been proposed to study uncertain RDF and XML databases. This paper discusses the reachability queries over uncertain graphs, specifically a probabilistic reachability (PR) query over an uncertain graph using the possible world semantics. It is proved that processing PR query is a #P-complete problem. The authors first propose a basic random algorithm to efficiently estimate the reachable probability with a high quality. To further improve the basic method, the authors introduce conditional distribution in random algorithm called conditional random algorithm (CRA), and compute the disjoint path set and cut set probabilities for the conditional distribution that is used in CRA, which helps us to find the querying results in polynomial time. Finally, the authors have verified the effectiveness of the proposed solutions for PR queries through extensive experiments on real uncertain graph datasets.

Keywords uncertain graph; possible world; conditional random algorithm; path set; cut set

收稿日期: 2010-06-11. 本课题得到国家自然科学基金重点项目(60933001)、国家自然科学基金面上项目(60773221)、国家"八六三"高技术研究发展计划项目基金(2009AA01Z150)、国家自然科学基金(60803026)资助. 袁 野, 男, 1981 年生, 博士研究生, 研究方向为图数据库、概率数据库、P2P 数据管理和数据隐私. E-mail: linuxyy@gmail.com. 王国仁, 男, 1966 年生, 教授, 博士生导师, 研究领域为 XML 数据管理技术、查询处理与优化、概率数据库和生物信息学.

1 引　言

本文研究一个不确定图上任意两点的可达性问题. 具体地, 给出一不确定图和其任意两点 s 和 d, 欲返回 s 和 d 的连通概率. 由于不确定图存在于许多应用中[1-5], 研究不确定图中的可达查询十分必要, 下面给出两个具体应用的实例.

应用实例 **1**.　文献[3]给出了一个音乐推荐社会网络. 该网络的每个节点代表一个音乐用户, 赋予每条边一概率来表示一个用户的音乐建议将被另一个用户采纳的可能性. 图 1 给出该网络的一部分, 在图中, 边 e_3 上的概率表示用户 B 的建议影响用户 A 的概率是 0.75. 对图 1, 会提出查询"用户 s 的建议影响用户 d 的概率是多少?".

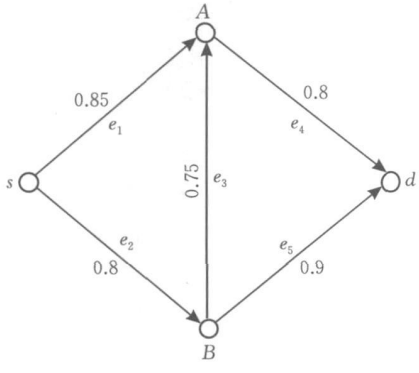


图 1　一个不确定图

应用实例 **2**.　在生物信息学中, 常用蛋白质作用(PPI)网络分析蛋白质之间的关系. 在 PPI 网络中, 节点表示蛋白质, 边表示两种蛋白质之间的相互作用, 其中可达查询是一种基本分析方法. 例如, 会提出"是否一种基因会间接地受另一种基因控制?或两种蛋白质之间是否存在生物路径?"等问题. 由于不准确的 PPI 检测方法, 使用不确定图表示 PPI 网络是合适的, 其中不确定边的权值表示蛋白质之间存在相互作用的可能性[4,6]. 文献[7]给出一个真实 PPI 网络, 其中每条边被赋予一个概率. 此概率被定义为两种蛋白质之间的可靠性, 该值越大, 两种蛋白质之间越可靠. 对于该网络, 可以用可达查询计算一种蛋白质到另一种蛋白质的可靠性.

从以上应用实例可看出, 基于不确定有向图的可达性查询比在传统确定有向图的可达查询能表达出更丰富的语义. 例如图 1 所示的音乐建议网络, 边的方向表示建议的方向. 传统的 s 和 d 之间的可达性查询仅返回 s 的建议可以影响 d. 但是, 例 1 中提出查询不仅返回"可以影响", 而且还返回影响的程度.

为解决不确定有向图的概率可达查询问题, 本文采用可能世界模型[8-9](一种被用来描述概率数据库的模型)表达不确定图. 具体地, 给出一个不确定有向图, 赋予每边一个存在概率, 根据文献[1-4, 7]应用的规定, 本文假设不同边的分布是独立的. 一个不确定图的可能世界是一个确定图, 称为可能世界图(简称可能图), 它是不确定图中所有边(取决于它们的存在性)组合的一个实例. 一个可能图的概率是其所有存在边概率和其不存在边的不存在概率的乘积. 给定不确定图两顶点 s 和 d, 从 s 到达 d 的可达概率是部分可能图概率的和, 在这些可能图中, s 和 d 必须是连通的.

图 1 给出一不确定图, 计算从 s 到 d 的可达概率. 很明显该不确定图可派生出 $2^5 = 32$ 个可能图. 表 1 列出了所有 s 到 d 连通的可能图和相应的概率. 那么 s 到 d 的可达概率即是这些可能图概率的和, 其值是 0.9176.

表 **1**　图 **1** 中 s 和 d 可达的可能世界

可能世界及其概率	概率的和(s 到 d 的可达概率)
$e_1e_2e_3e_4e_5(Pr=0.3672)$, $e_1\overline{e_2}e_3e_4e_5(Pr=0.0918)$, $e_1e_2\overline{e_3}e_4e_5(Pr=0.1224)$, $e_1e_2e_3e_4\overline{e_5}(Pr=0.0408)$ $e_1\overline{e_2}\overline{e_3}e_4e_5(Pr=0.0306)$, $e_1e_2\overline{e_3}e_4\overline{e_5}(Pr=0.0136)$, $e_1\overline{e_2}e_3e_4\overline{e_5}(Pr=0.0102)$, $e_1\overline{e_2}e_3e_4\overline{e_5}(Pr=0.0034)$ $\overline{e_1}e_2e_3e_4e_5(Pr=0.0648)$, $\overline{e_1}e_2e_3e_4\overline{e_5}(Pr=0.0072)$, $e_1e_2e_3\overline{e_4}e_5(Pr=0.0918)$, $\overline{e_1}e_2e_3\overline{e_4}e_5(Pr=0.0162)$ $e_1e_2\overline{e_3}\overline{e_4}e_5(Pr=0.0306)$, $\overline{e_1}e_2\overline{e_3}\overline{e_4}e_5(Pr=0.0054)$, $\overline{e_1}e_2\overline{e_3}e_4e_5(Pr=0.0216)$	0.9176

一种直接求解 PR 查询的方法是枚举不确定图所有的可能图, 并且对每个可能图做传统的可达查询处理. 找出所有给定两点连通的可能图, 对其概率求和, 所得结果即是可达概率. 该方法被称为 Naive 算法. 但是, 此方法的效率非常低, 因为需枚举指数级的可能图. 因此, 欲快速求解查询, 本文首先给出一基本随机算法, 该算法用于模拟产生可能图的过程, 从而可以快速在此过程中求解查询, 并用相关理论保证结果的准确性. 此算法的缺点是对较小的可达概率收敛速度慢, 而小可达概率是图数据库应用的主要特点, 比如 PPI 网络中, 蛋白质相互作用及其微弱. 为弥补基本随机算法的不足, 本文设计了

“条件随机算法”. 改进后的随机算法能在多项式时间内高精度地求解小概率可达查询.

2 问题定义

定义 **1**. 一个不确定图是集合 $G=((V,E),Pr))$, 其中(V,E)是有向无环图, $Pr: E^{\rightarrow}(0,1]$是定义边集中每条边存在的概率函数.

从定义 1 易知确定图是一个特殊的边概率为 1 的不确定图. 正如第 1 节提到的, 本文使用可能世界模型来定义不确定图. 在可能世界模型下, 一不确定图可派生出一组确定图 $G'=(V',E')$, 此确定图称为可能世界图, 简称可能图, 它满足 $V'=V, E\subseteq E'$.

根据文献[1-4, 7]中的规定, 本文假设不确定图不同边的概率分布是相互独立的, 因此可能图的概率为

$$Pr(G')=\prod_{e\in E'}Pr(e)\cdot\prod_{e\in E\setminus E'}(1-Pr(e)) \quad (1)$$

设 $Imp(G)$ 为不确定图 G 的所有可能图的集合. 显然 $Imp(G)$ 的大小是 $2^{|E|}$, 并且对任何可能图 G' 有 $Pr(G')>0$, $\sum Pr(G')=1$.

定义 **2**. 给一不确定图 G 和其两个顶点 s 和 d, 概率可达查询返回 s 到 d 的可达概率:

$$q_{pr}(s,d)=\sum_{G'\in RA(G,s,d)}Pr(G') \quad (2)$$

其中 $RA(G,s,d)$ 是 s 可达 d 的可能图的集合.

注意到 s 和 d 在 G 中可能不连通, 本文采用“Path-Tree Cover”[10]方法首先测试 s 是否可到达 d. 如果它们连通, 则计算可达性概率 q_{pr}. 否则, 直接返回不可达信息. 接下来, 本文都假设在 G 中 s 和 d 是连通的.

在给出具体 PR 查询复杂度前, 先给出如下概念.

定义 **3**. 给定一个不确定图 G 和其两个顶点 s 和 d, 一个 $s-d$ 割是一组边的集合, 删除该边集, s 和 d 不再连通.

如图 1 所示, $\{e_1,e_2\}$是一个 $s-d$ 割.

定理 **1**. 计算不确定图中的可达概率是一个 #P 完全问题.

证明. 假设 G 的 m 条边中, 每条边的不存在概率是 $1-p$ 那么 $q_{pr}=1-\sum_{i=0}^{m}\lambda(1-p)^i p^{m-i}$, 其中, λ是一个 $s-d$ 割的大小. 变换上式得

$$\sum_{i=0}^{m}\lambda\left(\frac{1-p}{p}\right)^i=p^{-m}(1-q_{pr}) \quad (3)$$

下面用反证法证明定理, 假设可在多项式时间内计算 q_{pr}. 将 $m+1$ 个不同 p 值代入方程(3), 得 $m+1$ 个线性方程, 其中有 $m+1$ 个未知数 λ和 $m+1$ 个已知数 p 和 q_{pr}. 该线性方程组的系数矩阵是非奇异的, 因此可用高斯消元迭代算法在多项式时间内求解方程组, 从而可以在解集中快速找到最小 $s-d$ 割的基数. 因此, 如果可以多项式时间计算 q_{pr}, 那么可多项式时间得到图 G 最小 $s-d$ 割的基数. 但是计算一图 $s-d$ 最小割的基数是一个 #P 完成问题[12], 从而推出矛盾. 证毕.

3 基本随机算法

从定理 1 知, 处理 PR 查询是 #P 完全问题, 这也意味着不存在快速算法计算精确的 q_{pr}. 因此, 本节给出一简单而有效的抽样方法近似计算 q_{pr}.

设 X_e是一随机变量, 以表示 e 存在于不确定图 G 派生出的可能图 G' 的事件, 如果 $e\in G'$, $X_e=1$, 否则, $X_e=0$. 令 $m=|E|$, 则随机向量 $\boldsymbol{X}=(X_1,\cdots,X_m)$可表示 G 派生出 G' 这一事件, 易得出 X 的状态个数为 2^m.

定义 $\Phi(\boldsymbol{X})$为

$$\Phi(\boldsymbol{X})=\begin{cases}1, & \text{如果 } s \text{ 到 } d \text{ 可达,}\\ 0, & \text{否则.}\end{cases}$$

根据式(2), 可得 q_{pr} 的值是 $E(\Phi(\boldsymbol{X}))$, 算法 1 给出近似求解 $E(\Phi(\boldsymbol{X}))$的步骤.

算法 **1**. *ApproxCompute()*.

1. $V=0$;
2. Repeat;
3. 对每边 $e\in E$, 以概率 $Pr(e)$ 选取“1”作为 X_e的值, 或以概率 $1-Pr(e)$ 选取“0”作为 X_e的值, 这样便可构成向量 $\boldsymbol{X}=(X_1,\cdots,X_m)$;
4. if $(\Phi(\boldsymbol{X})==1)$
5. $V=V+1$;
6. Until N times;
7. Return V/N.

当 N 的值足够大, Chernoff-Hoeffding 界给出近似解的精确保证[11].

引理 **1**. 对任意的 $\delta(0<\delta<1)$和 $\tau(\tau>0)$, 如果 $N\geq\left(3\ln\frac{2}{\delta}\right)\Big\backslash\tau^2$, 则 $Pr\{|V/N-E(\Phi(X))|>\tau E(\Phi(X))\}\leq\delta$

定理 2 保证随机算法 1 的近似解有很高的精确度.

4　改进的随机算法

本节为随机算法引入条件分布(称为"条件随机算法"(CRA))来改进基本算法. 改进的随机算法能在 $m(m=|E|)$ 多项式时间内完成查询, 同时得出高精度的结果.

本质上说, 算法 1 生成 $\boldsymbol{X}_1, \cdots, \boldsymbol{X}_N$ 与 $\boldsymbol{X}$ 同分布的 N 个独立随机向量, 得到的近似值 $\hat{q}_{BC}$ 可表示为

$$\hat{q}_{BC}=\frac{1}{N}\sum_{i=1}^{N}\Phi(\boldsymbol{X}_i) \tag{4}$$

这一近似结果的质量由 $100(1-\alpha)\%$ 置信区间所给出[1]:

$$\left[\hat{q}_{BC}-z_{1-\alpha/2}\frac{\hat{\sigma}_N}{\sqrt{N}},\ \hat{q}_{BC}+z_{1-\alpha/2}\frac{\hat{\sigma}_N}{\sqrt{N}}\right],$$

其中 $\hat{\sigma}_N$ 是基于 $\boldsymbol{X}_1, \cdots, \boldsymbol{X}_N$ 对 $Var(\Phi(\boldsymbol{X}))$ 的估计.

通常用置信区间的半宽度(HW)来衡量近似结果的质量. 对于固定的 α, 有

$$HW=z_{1-\alpha/2}\frac{Var(\Phi(\boldsymbol{X}))}{\sqrt{N}}=z_{1-\alpha/2}Var(\hat{q}_{BC}) \tag{5}$$

如果 HW 很小, 则说明估计的质量很高. 此外, 对较小的 q_{pr}, 采用公式(4)估计出的 $\hat{q}_{BC}$ 方差会很高, 有必要增加 N 来改善估计, 会导致 N 的值非常大. 相反如果控制 N 较小, 估计的 $\hat{q}_{BC}$ 误差就很大. 对于不确定图, 小的 q_{pr} 值可能出现在图中两个顶点距离非常大, 或者每条边的概率值非常小的情况下. 因此, 需要降低 $Var(\hat{q}_{BC})$ 以改善算法 1, 以保证对较大和较小的 q_{pr} 都能快速收敛. 一个有效降低 $Var(\hat{q}_{BC})$ 的方法是引入条件分布[13].

4.1　条件随机算法

首先介绍 CRA 的基本方法[13].

设 $C=C(\boldsymbol{X})$ 是一离散随机函数, 取值 $\{c_1, \cdots, c_k\}$. 又引入

$$\theta_i=E[\Phi\mid C=c_i],\quad i=1,2,\cdots,k \tag{6}$$

因此得

$$q_{pr}=\sum_{i=1}^{k}\theta_i Pr(C=c_i) \tag{7}$$

这里把集合 C 划分成 k 组, 每组对应一个 θ_i, 其中第 i 组的大小为 $N_i(i=1,2,\cdots,k)$, 并且 $N_1+\cdots+N_k=N$. 第 i 组中的取值是根据 $\boldsymbol{X}$ 的 $C=c_i$ 的条件分布生成的, 用来估计条件期望 $\theta_i(i=1,2,\cdots,k)$. 设 $\{\boldsymbol{X}_{r,i}: r=1,2,\cdots,N_i\}$ 表示第 i 组的取值, 则 θ_i 被估计为

$$\hat{\theta}_i=\frac{1}{N_i}\sum_{r=1}^{N_i}\Phi(\boldsymbol{X}_{r,i}),\quad i=1,2,\cdots,k \tag{8}$$

整合此 k 个估计, 可得 q_{pr} 的 CRA 估计:

$$\hat{q}_{CRA}=\sum_{i=1}^{k}\hat{\theta}_i Pr(C=c_i) \tag{9}$$

其中,

$$Var(\hat{q}_{CRA})=\sum_{i=1}^{k}Var(\Phi\mid C=c_i)[Pr(C=c_i)]^2/N_i \tag{10}$$

从式(10)可看出, 条件估计的方差依赖于 N_i. 理想情况下, 如果条件方差较大, 则希望 N_i 也较大, 反之亦然. 因此最佳的分组能保证最小的总方差. 然而, 实际中这是很困难的, 因为条件方差是未知的. 为和 $Var(\hat{q}_{BC})$ 进行比较, 这里取 $N_i\approx N\cdot Pr(C=c_i)$, $i=1,2,\cdots,k$, 那么,

$$\begin{aligned}Var(\hat{q}_{CRA})&\approx\sum_{i=1}^{k}Var(\Phi\mid C=c_i)Pr(C=c_i)/N\\&=E[Var(\Phi\mid C)]/N\\&=\{Var(\Phi)-Var[E(\Phi\mid C)]\}/N\\&\le Var(\Phi)/N=Var(\hat{q}_{BC})\end{aligned} \tag{11}$$

从式(11)可看出, CRA 估计比算法 1 所得估计有更小的方差. 这个数量可以解释为: C 包含多少与 Φ 相关的信息. 因此欲选择较好的 C, 应该选取含有尽可能多关于 Φ 信息的离散变量. 但是, 在应用 CRA 时还有其它一些重要的方面需要考虑[13]. 首先, C 必须能在多项式时间内计算出其分布. 其次, C 可能取值个数(这里指 k)不能超过 m 的多项式. 最后, 需可以对 X 关于 C 的条件分布进行快速抽样. 在下一小节中, 将给出满足上述条件的 C.

4.2　条件分布的上界和下界

本节给出具体应用 CRA 估算 q_{pr} 的算法, 它的基本框架如下:

首先选取式(7)中, $C(\boldsymbol{X})\triangleq\Phi_L(\boldsymbol{X})\le\Phi(\boldsymbol{X})\le\Phi_U(\boldsymbol{X})$ 作为 CRA 中的条件. 使用这些界限, 可得由式(12)给出的 q_{pr} 的估计值, 但这里需要产生独立同分布于 $Pr(\Phi(\boldsymbol{X})\mid C(\boldsymbol{X}))$ 的 N 个样本. 因此, 最重要的一步是如何计算条件分布 $Pr(\Phi(\boldsymbol{X})\mid C(\boldsymbol{X}))$, 设 S 表示该条件分布, 式(14)给出 S 的计算公式. 本节把 S 的计算转换为计算基于每边随机变量 X_i, $1\le i\le|E|$ 的条件分布, 其结果由式(15)给出. 选择不确定图 G 的不相交的路径集和割集向量作为 $\Phi(\boldsymbol{X})$ 的上下界. 结合这些界的性质, 给出基于迭代计算的

[1] Martin Haugh. http://www.columbia.edu/mh2078/MCS04.html

策略求解 X_i 的条件分布, 式(32)给出此结果. 利用这些结果, 便可计算条件分布 S. 最后, 我们可以使用式(12)来估计 q_{pr}.

要应用 CRA, 本文定义 $C(\boldsymbol{X}) \Leftrightarrow \phi_L(\boldsymbol{X}) \leq \phi(\boldsymbol{X}) \leq \phi_U(\boldsymbol{X})$, 其中 ϕ_L 和 ϕ_U 是不确定图对向量 $\boldsymbol{X}$ 的结构函数. 如果这两个函数非常逼近于 $\phi(\boldsymbol{X})$, 便可估算一个准确的 q_{pr}. 根据 C 的定义, C 的可能取值是 $\{(\phi_L, \phi_U)\} = \{(0,0), (0,1), (1,1)\}$, 又有 $E[\phi \mid \phi_L = \phi_U = 0] = 0$ 和 $E[\phi \mid \phi_L = \phi_U = 1] = 1$. 因此只需估计 $E[\phi \mid \phi_L = 0, \phi_U = 1]$. 观察到,

$$\begin{aligned} Pr(C=(0,1)) &= Pr(\phi_L = 0 \cap \phi_U = 1) \\ &= Pr(\phi_U = 0) - Pr(\phi_L = 1) \\ &= E[\phi_U] - E[\phi_L] = q_U - q_L, \end{aligned}$$

其中 q_U 和 q_L 分别是 $q_{pr} = E[\phi]$ 的上界和下界. 类似地可得

$$\begin{aligned} Pr(C=(1,1)) &= Pr(\phi_L = 1 \cap \phi_U = 1) \\ &= Pr(\phi_L = 1) = E[\phi_L] = q_L. \end{aligned}$$

将这些结果代入式(9), 得 q_{pr} 的 CRA 估计:

$$\hat{q}_{\mathrm{CRA}} = (q_U - q_L)\frac{1}{N}\sum_{r=1}^{N}\phi(\boldsymbol{X}_r) + q_L \tag{12}$$

其中 $\boldsymbol{X}_1, \cdots, \boldsymbol{X}_N$ 是 $\boldsymbol{X}$ 对 $(\phi_L = 0, \phi_U = 1)$ 的条件分布的样本. 设 S 表示此条件分布, 得估计的方差:

$$\begin{aligned} Var_S(\hat{q}_{\mathrm{CRA}}) &= \frac{1}{N}(q_U - q_L)^2 Var_S[\phi(\boldsymbol{X})] \\ &= \frac{1}{N}(q_U - q_L)^2 [E_S\phi(\boldsymbol{X}) - (E_S\phi(\boldsymbol{X}))^2] \\ &= \frac{1}{N}(q_U - q_L)^2 \left[\frac{q_{pr} - q_L}{q_U - q_L} - \left(\frac{q_{pr} - q_L}{q_U - q_L}\right)^2\right] \\ &= \frac{1}{N}(q_U - q_{pr})(q_{pr} - q_L), \end{aligned}$$

其中 E_S 和 Var_S 是 S 分布的期望和方差. 从上式可看出, 当 $q_{pr} = (q_U + q_L)/2$ 时, $Var_S(\hat{q}_{\mathrm{CRA}})$ 取最大值:

$$\max\{Var_S\} = \frac{(q_U - q_L)^2}{4N} \tag{13}$$

从式(13)易知, 当 q_U 和 q_L 尽可能接近时, 可得很小的方差, 这也意味着估计的收敛速度更快.

现在我们对条件分布进行采样. 首先引进了边状态概率 $P_i = Pr(X_i = 1)$, $i = 1, 2, \cdots, m$. 此时条件分布 S 记为

$$Pr(X_1 = x_1, \cdots, X_m = x_m \mid \phi_U(\boldsymbol{X}) - \phi_L(\boldsymbol{X}) = 1) = \prod_{i=1}^{m} Pr\left(X_i = x_i \mid \phi_U(\boldsymbol{X}) - \phi_L(\boldsymbol{X}) = 1, \bigcap_{k=1}^{i-1} X_k = x_k\right) \tag{14}$$

这样把对 S 的采样问题转化为对边随机变量 X_i 的条件分布的采样. 由于边的概率相互独立, 式(14)可写成

$$\begin{aligned} &Pr\left(X_i = x_i \mid \phi_U(\boldsymbol{X}) - \phi_L(\boldsymbol{X}) = 1, \bigcap_{k=1}^{i-1} X_k\right) \\ &= \frac{Pr\left(\phi_U(\boldsymbol{X}) - \phi_L(\boldsymbol{X}) = 1 \mid \bigcap_{k=1}^{i} X_k = x_k\right)}{Pr\left(\phi_U(\boldsymbol{X}) - \phi_L(\boldsymbol{X}) = 1 \mid \bigcap_{k=1}^{i-1} X_k = x_k\right)} \cdot \\ &\quad Pr\left(X_i = x_i \mid \bigcap_{k=1}^{i-1} X_k = x_k\right) \\ &= \frac{E\left(\phi_U \mid \bigcap_{k=1}^{i} X_k = x_k\right) - E\left(\phi_L \mid \bigcap_{k=1}^{i} X_k = x_k\right)}{E\left(\phi_U \mid \bigcap_{k=1}^{i-1} X_k = x_k\right) - E\left(\phi_L \mid \bigcap_{k=1}^{i-1} X_k = x_k\right)} Pr(X_i = x_i) \\ &= \frac{q_U(\boldsymbol{X}_i, \boldsymbol{P}) - q_L(\boldsymbol{X}_i, \boldsymbol{P})}{q_U(\boldsymbol{X}_{i-1}, \boldsymbol{P}) - q_L(\boldsymbol{X}_{i-1}, \boldsymbol{P})} p_i^{x_i}(1 - p_i)^{1 - x_i} \end{aligned} \tag{15}$$

其中 $q_U(\boldsymbol{X}_i, \boldsymbol{P}) = E\left(\phi_U \mid \bigcap_{k=1}^{i} X_k = x_k\right)$, $q_L(\boldsymbol{X}_i, \boldsymbol{P}) = E\left(\phi_L \mid \bigcap_{k=1}^{i} X_k = x_k\right)$.

从式(15)可见, 如果 ϕ_L 和 ϕ_U 可在多项式时间内完成计算, 那么可在多项式时间内完成对条件分布 S 的采样. 下面提出一种获得此上界和下界函数的方法.

本文使用不确定图 G 中不相交路径集和割集作为 $\phi(\boldsymbol{X})$ 的界限.

定义 **4**. 给定一个不确定图 G 和其两顶点 s 和 d, 一条 $s-d$ 路径是连通 s 和 d 的边集, 一个 $s-d$ 割是一边集, 删除这一边集, s 和 d 不再连通.

如图 1 所示, $\{e_1, e_4\}$ 和 $\{e_2, e_5\}$ 是两条路径, $\{e_1, e_2\}$ 和 $\{e_4, e_5\}$ 是两个割.

定义 PS 和 CS 分别为 G 的所有 $s-d$ 路径和割的集合. 这两个集合可唯一地确定 G, 因此可用来计算 q_{pr}. 从概率理论, 易得

$$\begin{aligned} q_{pr} &= E\left[1 - \prod_{Path \in PS}\left(1 - \prod_{i \in Path} X_i\right)\right] \\ &= E\left[\prod_{Cut \in CS} 1 - \prod_{i \in Cut}(1 - X_i)\right]. \end{aligned}$$

为简化表述, 上述公式改写为

$$q_{pr} = E\left[\coprod_{Path \in PS}\prod_{i \in Path} X_i\right] = E\left[\prod_{Cut \in CS}\coprod_{i \in Cut} X_i\right] \tag{16}$$

其中 $\coprod_i X_i = 1 - \prod_i (1 - X_i)$.

但计算式(16)是困难的, 因为路径集或割集里的元素通常不会两两不相交, 并且 PS 和 CS 的大小在最坏的情况下会以 m 的指数级增长. 设 PS' 和 CS' 分别是 PS 和 CS 的子集, 只由不相交的路径和割组成. 已有具体算法可以直接产生它们, 而不需要产生集合 PS 和 CS. 此外, PS' 和 CS' 的大小总是小于 m.

定义 $\Phi_L(\boldsymbol{X})$ 和 $\Phi_U(\boldsymbol{X})$ 如下,

$$\Phi_L(\boldsymbol{X})=\coprod_{Path\in PS'}\prod_{i\in Path}X_i,\quad \Phi_U(\boldsymbol{X})=\prod_{Cut\in CS'}\coprod_{i\in Cut}X_i.$$

由于 PS' 和 ES' 中的元素不相交, 因此元素对应的事件是独立的, 为此有

$$\begin{cases}E[\Phi_L(\boldsymbol{X})]=\coprod\limits_{Path\in PS'}\prod\limits_{i\in Path}E[X_i],\\ E[\Phi_U(\boldsymbol{X})]=\prod\limits_{Cut\in CS'}\coprod\limits_{i\in Cut}E[X_i]\end{cases}\tag{17}$$

由于 PS' 和 CS' 分别是 PS 和 CS 的子集, 由式(16)和(17)可得

$$E[\Phi_L(\boldsymbol{X})]\le q_{pr}=E[\Phi(\boldsymbol{X})]\le E[\Phi_U(\boldsymbol{X})].$$

从上式可见, 如果 PS' 和 CS' 的基数尽可能大, 就会得到更紧的界. 为此本文应用网络流理论中的两个定理[14].

引理 **2**.　最大不相交 $s-d$ 路径数等于最小 $s-d$ 割的基数.

引理 **3**.　最大不相交 $s-d$ 割数等于最短 $s-d$ 路径的长度.

这里, 我们使用文献[14]中的算法求解 PS' 和 CS' 基数的最大值, 该算法可在多项式时间内完成. 在图 1 中, PS' 和 CS' 的最大集合分别为 $\{\{e_1,e_4\},\{e_2,e_5\}\}$ 和 $\{\{e_1,e_2\},\{e_4,e_5\}\}$. 因此 $E[\Phi_L]=1-(1-P_1P_4)(1-P_2P_5)=1-(1-0.85\times0.8)(1-0.8\times0.9)=0.9104$, $E[\Phi_U]=1-(1-P_1P_4)(1-P_2P_5)=0.9506$. 由于 $PS'\in PS$, $CS'\in CS$, 因此有 $\Phi_L\le\Phi\le\Phi_U$. 例如, 在图 1 中, $(\Phi_L=0.9104)\le(\Phi=0.9176)\le(\Phi_U=0.9506)$.

因此, 式(15)中计算分布 S 所需的条件期望可计算如下:

$$E(\Phi_L\mid\bigcap_{k=1}^{r}X_k=x_k)=q_L(\boldsymbol{X}_r,\boldsymbol{P})=\coprod_{Path\in PS'}\Big[\prod_{k\in Path\setminus E_r}p_k\Big]\cdot\Big[\prod_{k\in Path\cap E_r}x_k\Big],$$

$$E(\Phi_U\mid\bigcap_{k=1}^{r}X_k=x_k)=q_U(\boldsymbol{X}_r,\boldsymbol{P})=\prod_{Cut\in CS'}\Big[\coprod_{k\in Cut\setminus E_r}p_k\Big]\cdot\Big[\coprod_{k\in Cut\cap E_r}x_k\Big],$$

其中 $E_r=\{1,2,\cdots,r\}$, $r=1,2,\cdots,m$.

对式(15)所需剩余的计算项, 我们构造递推式来计算. 为此引入如下符号,

$$\begin{cases}a_{Path}(X_r,P)=\Big[\prod\limits_{k\in Path\setminus E_r}p_k\Big]\Big[\prod\limits_{k\in Path\cap E_r}x_k\Big],\\ \qquad Path\in PS',\ r=1,2,\cdots,m\\ b_{Cut}(X_r,P)=\Big[\coprod\limits_{k\in Cut\setminus E_r}p_k\Big]\Big[\coprod\limits_{k\in Cut\cap E_r}x_k\Big],\\ \qquad Cut\in CS',\ r=1,2,\cdots,m\end{cases}\tag{18}$$

使用上述符号, 可重写 $q_L(\boldsymbol{X}_r,\boldsymbol{P})$ 和 $q_U(\boldsymbol{X}_r,\boldsymbol{P})$ 为

$$\begin{cases}q_L(\boldsymbol{X}_r,\boldsymbol{P})=\coprod\limits_{Path\in PS'}a_{Path}(\boldsymbol{X}_r,\boldsymbol{P}),\\ q_U(\boldsymbol{X}_r,\boldsymbol{P})=\coprod\limits_{Cut\in CS'}b_{Cut}(\boldsymbol{X}_r,\boldsymbol{P})\end{cases}\tag{19}$$

之后, 假设 $q_L(\boldsymbol{X}_{i-1},\boldsymbol{P})$ 和 $q_U(\boldsymbol{X}_{i-1},\boldsymbol{P})$ 已算好, 并且要计算 $q_L(\boldsymbol{X}_i,\boldsymbol{P})$ 和 $q_U(\boldsymbol{X}_i,\boldsymbol{P})$. 由于 PS' 中的元素不相交, 最多存在一个路径 $path^*\in PS'$, 使得 $i\in path^*$. 同样最多存在一个割 $cut^*\in CS'$, 使得 $i\in cut^*$. 此时由式(14)可得

$$\begin{cases}a_{Path^*}(\boldsymbol{X}_i,\boldsymbol{P})=\dfrac{x_i}{p_i}a_{Path^*}(\boldsymbol{X}_{i-1},\boldsymbol{P})\\ b_{Cut^*}(\boldsymbol{X}_i,\boldsymbol{P})=1-\dfrac{1-x_i}{1-p_i}[1-b_{Cut^*}(\boldsymbol{X}_{i-1},\boldsymbol{P})]\end{cases}\tag{20}$$

对从 $i-1$ 到 i 没起作用的路径和割有

$$\begin{aligned}&a_{Path}(\boldsymbol{X}_i,\boldsymbol{P})=a_{Path}(\boldsymbol{X}_{i-1},\boldsymbol{p}),\\ &\forall Path\in Path',\ Path\ne Path^*\\ &b_{Cut}(\boldsymbol{X}_i,\boldsymbol{P})=b_{Cut}(\boldsymbol{X}_{i-1},\boldsymbol{p}),\\ &\forall Cut\in Cut',\ Cut\ne Cut^*\end{aligned}\tag{21}$$

将式(20)和(21)代入式(19), 得最终结果:

$$q_L(\boldsymbol{X}_i,\boldsymbol{P})=1-\frac{1-a_{Path}\times(\boldsymbol{X}_i,\boldsymbol{P})}{1-a_{Path}\times(\boldsymbol{X}_{i-1},\boldsymbol{P})}[1-q_L(\boldsymbol{X}_{i-1},\boldsymbol{P})],$$

$$q_U(\boldsymbol{X}_i,\boldsymbol{P})=\frac{b_{Cut}\times(\boldsymbol{X}_i,\boldsymbol{P})}{b_{Cut}\times(\boldsymbol{X}_{i-1},\boldsymbol{P})}q_U(\boldsymbol{X}_{i-1},\boldsymbol{P})\tag{22}$$

当 $q_U(\boldsymbol{X}_{i-1},\boldsymbol{P})$ 和 $q_L(\boldsymbol{X}_{i-1},\boldsymbol{P})$ 已知时, 用式(22)可以在常数时间内计算 $q_U(\boldsymbol{X}_i,\boldsymbol{P})$ 和 $q_L(\boldsymbol{X}_i,\boldsymbol{P})$. 然后将 $q_U(\boldsymbol{X}_i,\boldsymbol{P})$ 和 $q_L(\boldsymbol{X}_i,\boldsymbol{P})$ 的结果代入式(15), 其结果再输入式(14), 最后可用式(14)求解条件分布 S. 这个过程可在 $O(m)$ 时间完成计算. 因此, 一个采样过程(用式(12)产生的一个 $\boldsymbol{X}_i$)可在 $O(m)$ 时间内完成. 有了这些结果, 便可由式(12)来估计 q_{pr}. 采样的次数 N 仍采用定理 2 给出的值以保证结果的准确性.

5　性能分析

本节用真实和合成的不确定图数据验证本文的算法, 算法代码用 Visual C++ 6.0 编写, 运行环境是奔腾 4 3.0GHz CPU, 2G 内存和 160G 硬盘. 对本文给出的算法, R-BC 表示基本随机算法, R-CRA 表示"条件随机算法". 真实不确定数据采用 Entrez gene 6091-OMIM 127700 数据库[1]. 它是描述人类基因网络的数据库, 其中网络节点代表"显性基因", 边被赋予小数权值以表示基因相互作用的大小. 使用该数据库的 5 个基因网络作为不确定图数据, 表 2 给出它们的参数.

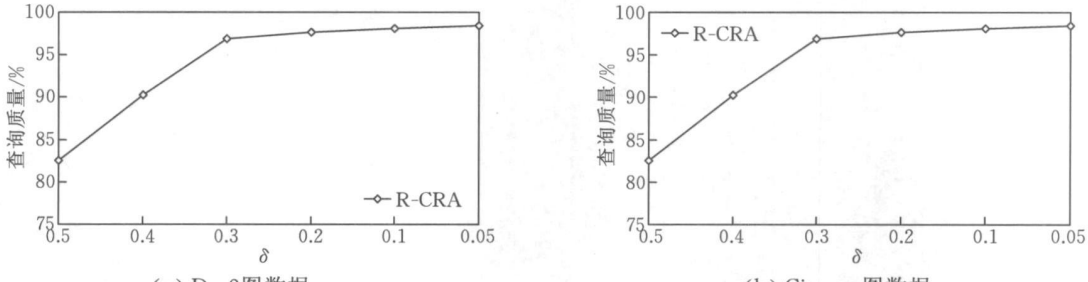
[1] http://www.ncbi.nlm.nih.gov/omim

表 2 真实不确定图数据的参数

图数据编号	节点数量	边集大小	边的平均概率
Dys1	452	1113	0.423
Dys2	1775	4163	0.396
Dys3	3259	8790	0.212
Dys4	6786	14056	0.237
Dys5	11368	32754	0.311

同时从 Citeseer[1] 中抽取图数据($|V|=12140$, $|E|=36874$), 并按正态分布 $N(0.7, 0.2)$ 为每边产生概率. 每次实验时, 随机地产生 100 个查询, 记录下查询的平均代价. 作为比较, 本文也实现了第 1 节给出的 Naive 算法, 因为目前还没有对此查询的工作.

正如前文所述, R-BC 和 R-CRA 的模拟次数都采用引理 1 给出的理论值, 该理论值是由 τ 和 δ 来控制的. 这两个参数是需要输入的, 为取得对实际应用的最佳参数值, 首先通过实验确定它们的取值. 固定 $\tau=0.1$, 改变 δ 的取值, 用模拟值和真实值 q_{pr} 的比来衡量不同 δ 取值下的查询质量. 因为 R-CRA 有更好的查询质量, 这里用 R-CRA 来测试. 这里通过测试真实数据和合成数据给出结果. Dys2 是小规模真实数据, Citeseer 是大规模的合成数据, 故选取它们更符合实际的应用. 图 2 给出对数据 Dys2 和 Citeseer 的测试结果. 从结果看出, 查询质量随着 δ 的减小而增加. 当 δ 减少到 0.3 以后, 曲线的增长变得非常平缓, 其中两组数据的查询质量都超过 95%. 当 $\delta=0.3$ 时, 由定理 2 得 $N=570$, 这是一个合理的模拟次数. 而对更小的 δ, 例如 $\delta=0.05$ 时, $N=1107$, 这需要相当长的时间, 而查询质量只比 $\delta=0.3$ 时略高些. 因此这里取$(\tau, \delta)=(0.1, 0.3)$.



(a) Dys2图数据 (b) Citeseer图数据

图 2 取不同 δ 时, 采样算法的查询质量

其次用真实数据测试算法的运行效率和可扩展性. 图 3 给出测试结果, 其中横坐标是 Dys1~ Dys5, 纵坐标是运行时间. 如图所示, 所有曲线都随图规模的增加而增长. 其中 Naïve 增长的最快, 到 Dys2 时就已经超过 100s, Naïve 要枚举所有可能图, 从而导致其时间的指数级增长. 而 R-CRA 和 R-BC 都具有较好的可扩展性, 即使对边数量超过 3 万的图 Dys5 求解 NP 难查询, 它们的运行效率也是高效的, 可在 1min 内完成. 从结果中也可发现, 尽管采用相同的模拟次数, R-CRA 比 R-BC 花费略多的时间, 因为 R-CRA 还需要计算图的路径集和割集.

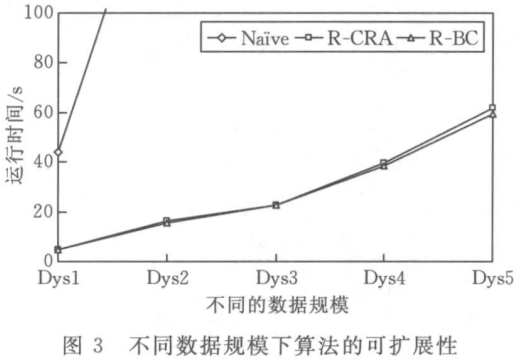


图 3 不同数据规模下算法的可扩展性

最后测试 R-BC 和 R-CRA 的查询质量. 正如第 4 节所述, 在运行效率相同时, R-CRA 有更高的查询质量, 尤其对较小的 q_{pr}. 因为真实数据的各参数已固定, 这里用合成数据来检验, 查询质量仍被定义为模拟值与真实值之比. 为模拟小 q_{pr}, 改变正态分布的 μ 从 0.8~ 0.05, 例如当 μ 取 0.05 时, q_{pr} 的数量级是 10^{-6}. 图 4 给出测试结果, 结果表明即使 q_{pr} 小至 10^{-6} 时, R-CRA 的查询质量也高于 90%. 而在 μ 取 0.4 和 0.8 时, 质量都在 98% 以上, 而此种情况更符合真实的应用[3-5]. 相反随着 μ 的减少, R-BC 的

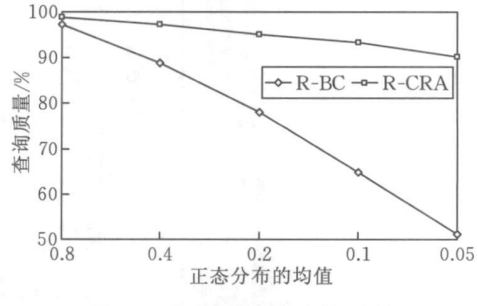


图 4 不同均值下的查询质量

[1] http://citeseer.ist.psu.edu/oai.html

质量急速下滑, 最小时已跌至 50%. 但在 μ大于 0. 4 时, R-BC 的查询质量也高于 90%, 这说明在实际应用中, R-BC 也有较可观的准确率. 图 5 给出在不同方差 σ 下查询的质量. 从图中可见, R-CRA 的曲线比 R-BC 稳定得多, 这说明 R-CRA 能较好地适应数据噪声动态变化的环境.

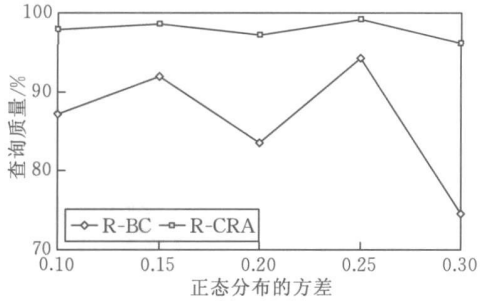


图 5　不同方差下的查询质量

6　相关工作

已有很多面向确定图的可达查询工作. 文献[15]用图的传递闭包来压缩原始图, 并在求解过程中把问题转化为网络流问题. 文献[16]把图分成若干有向链, 并每个顶点都记录了与其相邻的链, 从而可常数时间求解查询. 文献[17]首先提出树覆盖的方法, 并以树为单元对图进行压缩, 并证明此种压缩是保证可以求解查询的最优压缩. 文献[18]和文献[19]都是在索引构建时间上改进了树覆盖方法. 文献[10]拓展了有向链的方法, 把问题转换成平面图的问题, 并只需 2 跳即可完成查询.

面向概率数据库的研究是现在的热点. 早期工作的重点是在概率关系数据库上如文献[8-9, 20], 即处理概率 SQL 查询. 之后研究者提出了一些查询类型及其处理方法, 主要包括 Top-k 查询[21]和 Skyline 查询等[22]. 近年来开始关注结构化数据如概率 XML 数据[1, 2]. 就图数据库而言, 邹兆年等人[23]研究了不确定图的频繁子图挖掘, 而张硕等人[24]研究了带索引的不确定图 Top-k 查询.

7　结束语

本文所述是第一个面向不确定图研究可达查询的工作. 采用细粒度的可能世界模型定义不确定图, 从而使可达查询具有丰富的概率语义. 在给出问题是# P 难后, 本文首先设计了一个基本随机算法求解查询, 在理论分析了它的不足后, 采用"条件随机算法"改进基本算法. 其中选择不确定图的不相交路径集和割集作为可达概率的下界和上界去逼近真实概率, 计算基于该界的条件分布可在多项式时间内完成, 因此可多项式时间计算真实概率. 此外基本和改进的随机算法都保证了高质量的估计结果. 实验结果验证了本文的设计. 今后将研究基于不确定图的多种查询.

参　考　文　献

[1] Nierman A, Jagadish H V. ProTDB: Probabilistic data in XML//Proceedings of the VLDB. Hong Kong, China, 2002: 646-657

[2] Senellart P, Abiteboul S. On the complexity of managing probabilistic XML data//Proceedings of the PODS. Beijing, China, 2007: 283-289

[3] Adar E, Re C. Managing uncertainty in social networks. IEEE Data Engineering Bulletin, 2007, 30(2): 15-22

[4] Asthana S, King O D, Gibbons F D et al. Predicting protein complex membership using probabilistic network reliability. Genome Research, 2004, 14(6): 1170-1175

[5] Chui H N, Sung W K, Wong L. Exploiting indirect neighbors and topological weight to predict protein function from protein-protein interactions. Bioinformatics, 2007, 22(13): 47-58

[6] Jiang R, Tu Z, Chen T et al. Network motif identification in stochastic networks. PNAS, 2006, 103(25): 9404-9409

[7] Saito R, Suzuki H, Hayashizaki Y. Interaction generality: A measurement to assess the reliability of a protein-protein interaction. Nucleic Acids Research, 2002, 30(5): 1163-1168

[8] Dalvi N N, Suciu D. Management of probabilistic data: Foundations and challenges//Proceedings of PODS. Beijing, China, 2007: 1-12

[9] Khoussainova N, Balazinska M. Towards correcting input data errors probabilistically using integrity constraints//Proceedings of the MobiDE Workshop. Chicago, Illinois, USA, 2006: 43-50

[10] Jin R, Xiang Y, Ruan N. Efficiently answering reachability queries on very large directed graphs//Proceedings of SIGMOD. Vancouver, Canada, 2008: 595-608

[11] Angluin D, Valiant L G. Fast probabilistic algorithms for hamiltonian circuits and matchings//Proceedings of STOC. Boulder, Colorado, United States, 1977: 30-41

[12] Garey M R, Johnson D S. Computers and Intractability: A Guide to the Theory of NP-Completeness. New York: W. H. Freeman and Company, 1979

[13] Asmussen S, Glynn P. Stochastic Simulation: Algorithms and Analysis. New York: Springer, 2007

[14] Cormen T H, Leiserson C E, Rivest R L et al. Introduction to Algorithms. New York: The MIT Press/McGraw-Hill Book Company, 2001

[15] Jagadish H V. A compression technique to materialize transitive closure. TODS, 1990, 15(4): 558-598

[16] Cheng J, Yu J, Lin X. Fast computing reachability labelings for large graphs with high compression rate//Proceedings of the EDBT. Nantes, France, 2008: 193-204

[17] Agrawal R, Borgida A, Jagadish H V. Efficient management of transitive relationships in large data and knowledge bases. ACM SIGMOD Record, 1989, 18(2): 253-262

[18] Chen L, Gupta A, Kurul M. Stack-based algorithms for pattern matching on dags//Proceedings of the VLDB. Trondheim, Norway, 2005: 493-504

[19] Tribl S, Leser U. Fast and practical indexing and querying of very large graphs//Proceedings of the SIGMOD. Beijing, China, 2007: 845-856

[20] Benjelloun O, Sarma A D, Hayworth C. An introduction to ULDBs and the trio system. IEEE Data Engineering Bulletin, 2006, 29(1): 5-16

[21] Soliman M A, Ilyas I F, Chang K C. Top-*k* query processing in uncertain databases//Proceedings of ICDE. Istanbul, 2007: 896-905

[22] Pei J, Jiang B, Lin X. Probabilistic skylines on uncertain data//Proceedings of the VLDB. Vienna, Austria, 2007: 15-26

[23] Zou Zhao-Nian, Li Jian-Zhong, Gao Hong, Zhang Shuo. Ming frequent subgraph patterns from uncertain graphs. Journal of Software, 2009, 20(11): 2965-2976(in Chinese)
(邹兆年, 李建中, 高宏, 张硕. 从不确定图中挖掘频繁子图模式. 软件学报, 2009, 20(11): 2965-2976)

[24] Zhang Shuo, Gao Hong, Li Jian-Zhong, Zou Zhao-Nian. Efficient query processing on uncertain graph databases. Chinese Journal of Computers, 2009, 32(10): 2066-2079 (in Chinese)
(张硕, 高宏, 李建中, 邹兆年. 不确定图数据库中高效查询处理. 计算机学报, 2009, 32(10): 2066-2079)

YUNA Ye, born in 1981, Ph. D. candidate. His research interests include graph database, probabilistic database, P2P data management, and data piracy.

WANG Guo-Ren, born in 1966, professor, Ph. D. supervisor. His research interests include XML data management, query processing and optimization, probabilistic database, and bioinformatics.

Background

Efficiently answering reachability queries against very large graphs is becoming an increasingly important research topic driven by many emerging real world applications, such as biological networks, social networks, ontologies, XML and RDF databases. However, data extracted from those applications is inherently uncertain due to noise, incompleteness and inaccuracy, and many works have been proposed to study uncertain RDF and XML databases. Therefore, it is important and necessary to efficiently process reachability queries over graph data with uncertainty.

The topics on managing uncertain data are very hot nowadays, and there have been huge number of works on this topic. Initial works have focused on how to store and process uncertain data within database systems, and thus how to answer SQL-style queries. Subsequently, there has been a growing realization that in addition to storing and processing uncertain data such as KNN query, range query, top-k query and skyline query. There also exists several advanced algorithms to analyze uncertain data, e.g., clustering uncertain data and finding frequent items within uncertain data. For uncertain graph databases, the existing works have proposed algorithms for mining frequent subgraphs patterns and finding top-k patterns from uncertain graphs. This paper focuses on processing uncertain reachability queries that are not only common on uncertain graph databases, but also serve as fundamental operations for many other uncertain graph queries.

This research was supported by the National Natural Science Foundation of China (grant No. 60773221), the National High Technology Research and Development Program (863 Program) of China (grant No. 2009AA01Z150), and the National Basic Research Program(973 Program) of China (grant No. 60933001).