计算机研究与发展
Journal of Computer Research and Development
ISSN 1000-1239/CN 11-1777/TP
48(10): 1850-1858, 2011

不确定图上的 kNN 查询处理

张应龙[1,2] 李翠平[1] 陈 红[1] 杜凌霞[1]
[1](数据工程与知识工程教育部重点实验室(中国人民大学) 北京 100872)
[2](江西农业大学计算机与信息工程学院 南昌 330045)
(zhang_yinglong@126.com)

k-Nearest Neighbors in Uncertain Graph

Zhang Yinglong[1,2], Li Cuiping[1], Chen Hong[1], and Du Lingxia[1]
[1](*Key Laboratory of Data Engineering and Knowledge Engineering (Renmin University of China), Ministry of Education, Beijing* 100872)
[2](*School of Computer and Information Engineer, Jiangxi Agriculture University, Nanchang* 330045)

Abstract In many areas, a lot of data have been modeled by graphs which are subject to uncertainties, such as molecular compounds and protein interaction networks. While many real applications, for example, collaborative filtering, fraud detection, and link prediction in social networks etc, rely on efficiently answering k-nearest neighbor queries (kNN), which is the problem of computing the most "similar" k nodes to a given query node. To solve the problem, in this paper a novel method based on measurement of SimRank is proposed. However, because graphs evolve over time and are uncertainly, the computing cost can be very high in practice to solve the problem using the existing algorithms of SimRank. So the paper presents an optimization algorithm. Introducing path threshold, which is suitable in both determined graph and uncertain graph, the algorithm merely considers the local neighborhood of a given query node instead of whole graph to prune the search space. To further improving efficiency, the algorithm adopts sample technology in uncertain graph. At the same time, theory and experiments interpret and verify that the optimization algorithm is efficient and effective.

Key words uncertain graph; possible world; SimRank; k nearest neighbor; subgraph

摘 要 在现实中的许多领域产生大量不确定的图结构的数据,例如分子化合物、蛋白质交互网络等.同时现实中有很多应用例如推荐系统中的推荐过滤、欺诈检测和社会网络的链接预测等,需要查询给定节点的 k 个最相似节点,针对这一问题,提出了用基于 SimRank 度量的方法来求解.由于图的动态演变和不确定性导致用现有的 SimRank 计算方法求 k 个最近邻的代价昂贵,因此提出一个有效算法,在保证一定准确性的前提下,通过引入路径阈值,算法只需考虑查询点的邻居区域无需考虑整个图从而达到明显的剪枝效果,该方法在确定图和不确定图上都可以适用.在此基础上为了进一步提高效率,算法在不确定图上引入采样技术.最后从理论、实验说明验证了算法的高效性和有效性.

关键词 不确定图;可能世界;SimRank;kNN;子图

中图法分类号 TP311

收稿日期:2011-06-23;**修回日期**:2011-08-26

基金项目:国家自然科学基金项目(61070056,61033010);"核高基"国家科技重大专项基金项目(2010zx01042-001-002-002);教育部新世纪优秀人才支持计划基金项目(NECT-09-823)

通作作者:李翠平(cuiping_li@263.net)

从社会网络到蛋白质交互网络等不同领域产生的大量可用图结构表示的数据,由于获取数据的工具不精确或方法有误差等多种原因使得数据具有不确定性.在不确定图上最近邻查询在现实当中有很多应用,比如在蛋白质交互网络中需要查询"与 肌红蛋白最相近的 10 个蛋白质是哪些?",在论文的作者合作图中需要查找"与韩家炜教授研究最相似的 5 位研究者是哪几位?".因此有必要进行这方面的探究.

关于在不确定数据及图数据上的研究近年正在如火如荼地进行[1-8],研究点主要集中在不确定图上的频繁子图模式挖掘和计算最可靠子图.其中文献[1]和本文的研究最接近,文献[1]是在不确定图上基于距离的最近邻研究,然而文献[9]中的研究表明在图中基于距离来计算节点之间相似度的结果不尽人意,因此需要一种更准确的度量在不确定图上进行最近邻查询.

SimRank[10]是基于路径的节点相似度计算方法中最具有影响力的一种,迄今已经成功地应用到很多不同的领域.比如文献[11]将其应用到点击图中客户查询关键字与商业广告的匹配上,文献[12-13]将其应用到网络节点的聚类分析上.由于 SimRank 在计算节点相似度时,不但考虑了节点间的直接联系,而且考虑了节点间的间接联系,因此它的计算结果具有较高的准确性.

现有的 SimRank 计算方法针对所有节点采用全局循环迭代的方法实现的,所以当网络发生变化时,即便只有一个节点或一条边发生变化,要获得新的值,也必须对所有的节点进行重新计算.这种方法既低效,又无法适用于大数据量.文献[14]虽然可以对单个节点对计算出 SimRank 值,但在查询最近邻时代价也是昂贵的.因此在图动态变化情况下,现有方法是不能直接用在不确定图上进行最近邻查询的.

本文是在不确定图上进行基于 SimRank 的 k 个最近邻查询的研究.主要贡献如下:

1) 定义了不确定图上基于 SimRank 的 3 种度量;

2) 提出了一个有效的算法,可以同时在确定图和不确定图上求 kNN,该算法可以解决现有 SimRank 算法只考虑节点间长度为偶数的路径,从而导致在有些情况下所得出的节点相似度为零的缺陷;

3) 从理论上论证了算法的有效性和准确性;

4) 在真实的数据集上考察和证实了优化算法的可行性.

1 SimRank 简介

给定有向图 $G=\langle V,E\rangle$,$S(a,b)\in[0,1]$表示了图中节点 a 和 b 之间的 SimRank 相似度,计算公式为:

$$S(a,b)=\begin{cases}\dfrac{c}{|I(a)||I(b)|}\sum\limits_{i}^{|I(a)|}\sum\limits_{j}^{|I(b)|}S(I_i(a),\\ \quad I_j(b)),\ a\neq b;\\ 1,\ a=b;\end{cases}$$

$I(v)$表示图中节点 v 的入度邻居集合,$I_i(v)$为该集合中的一个元素.如果是无向图,则对应的是节点邻居集合.节点对节点本身的相似度是最大,例如 $S(a,a)=1$.当集合 $I(a)$或 $I(b)$为空集,避免零为除数,这时规定 $S(a,b)=0$.c 为衰减度因子.

令 $\boldsymbol{S}$ 为图 $\boldsymbol{G}$ 的相似度矩阵,迭代计算的矩阵形式为:

$$\boldsymbol{S}^k=\begin{cases}c\boldsymbol{W}^{\mathrm{T}}\boldsymbol{S}^{k-1}\boldsymbol{W}+(1-c)\boldsymbol{I},\ k>0;\\ \boldsymbol{I},\ k=0;\end{cases}$$

$\boldsymbol{W}$ 是图的邻接矩阵基于列上的标准化(column-normalized)矩阵.有关 SimRank 计算的详细信息可参考文献[10,15].

不失一般性,本文在没有具体说明时假设图是无向图.

2 问题描述

已知不确定图$\mathcal{G}=\langle V,E,P\rangle$,$V$,$E$ 分别表示图的顶点集合和边的集合,P 是与图中边关联的概率函数,$P(e)$表示边 e 存在的概率,这里 $e\in E$.本文假设不确定图中每条边的存在与否是相互独立的.图 G 是从不确定图$\mathcal{G}$中由概率 P 产生的确定图,即 E 中的每条边 e 以 $P(e)$的概率作为图 G 的边,这时称图 G 为不确定图的可能世界,记为 $G\sqsubseteq\mathcal{G}$.令 E_G 表示可能世界 G 的边的集合,则 G 存在的概率为 $Pr(G)=\prod\limits_{e\in E_G}p(e)\prod\limits_{e\in E-E_G}(1-p(e))$;令$\{G\}_P$ 为不确定图$\mathcal{G}$所有可能世界的集合,则集合$\{G\}_P$ 元素的个数为$2^{|E|}$.有关不确定图的相关信息可参考文献[1,6].

定义 1. a 与 b 无关.当两个节点 a,b 相似值小于给定阈值 α,即 $S(a,b)<\alpha$ 时,就可以认为节点 a

和 b 基本不相似,这时称 a 与 b 无关.

在不确定图$\mathcal{G}$中 $2^{|E|}$ 个可能世界中,每个可能世界都对应节点 a 和 b 的一个 SimRank 值.因此定义点 a 和 b 的 SimRank 的值分布为

$$p_{a,b}(s)=\sum_{G|S_G(a,b)=s\wedge\alpha\leqslant s<1}Pr[G],$$

这里 $G\sqsubseteq\mathcal{G}$,即 G 是$\mathcal{G}$的一个可能世界,$S_G(a,b)$ 表示点 a 和 b 在可能世界 G 中的 SimRank 值,α 为阈值;$p_{a,b}(s)$表示点 a 和 b 在不确定图上的 SimRank 值取 s 时的概率.

为了更准确地刻画不确定图中节点的相似度,定义不确定图中节点 a 和 b 的基于 SimRank 的 3 种度量:

这 3 种度量都是在不确定图$\mathcal{G}=\langle V,E,P\rangle$和 a,b 为其上任意两点的前提下定义的.

定义 2. a 和 b 在不确定图中的 SimRank 中位数.

$$S_M(a,b)=\arg\max_{s_k}\Big\{K\mid\sum_{t=0}^{k-1}p_{a,b}(s_t)<\frac{1}{2}\wedge\sum_{t=0}^{k}p_{a,b}(s_t)\geqslant\frac{1}{2}\Big\},$$

其中 s_t 为 a 和 b 在其中某个可能世界中一个 SimRank 值并且 $\alpha\leqslant s_t<s_{t+1}<1,t=0,1,2,\cdots$.

定义 3. a 和 b 在不确定图中的 SimRank 主值.

$$S_J(a,b)=\arg\max_s p_{a,b}(s).$$

定义 4. a 和 b 在不确定图中的 SimRank 期望值.

$$S_{ER}(a,b)=\sum s\times p_{a,b}(s).$$

文中把 SimRank 的中位数、主值、期望值分别简记为 S_M,S_J,S_{ER}.

不确定图$\mathcal{G}$节点相似度的最近邻查询是指给定一个查询点 a 和数值 k,返回和节点 a 最相似的 k 个节点.最近邻问题具体定义如下.

节点相似度的最近邻问题:在不确定图$\mathcal{G}=\langle V,E,P\rangle$上给定查询点 a、阈值 α、基于 SimRank 的概率度量 M(M 是 S_M,S_J 和 S_{ER}中任何一个度量)和 k,返回 k 个点的集合 $T_k(a)=\{t_1,t_2,\cdots,t_k\}$,其中集合中元素 t_i 满足:

$$M(a,t_i)\geqslant M(a,t)\ \forall t\in V\backslash T_k(a).$$

3 基本算法

首先对不确定图$\mathcal{G}$的每个可能世界,利用第 1 节中介绍的 SimRank 的迭代计算的矩阵公式计算相似度值,然后根据 3 种基于 SimRank 度量的定义,分别求出查询点和其他节点的基于这 3 种度量的值,由这些度量值确定查询点的 k 个最近邻点.基本算法如下:

算法 1. 基本算法.

输入:不确定图$\mathcal{G}$、查询点 a,k、阈值 α;

输出:点 a 的 k 个最近邻节点.

for($i=0;i<2^{|E|},i++$)

对每个可能世界 G 进行 SimRank 值计算;

求出 $Pr(G)$;

end for

由$M(S_M,S_J,S_{ER})$定义求出点 a 和其余点的度量值;

根据度量值求出 k 个最近邻.

基本算法的主要代价是对每个可能世界进行了 SimRank 计算.每个可能世界 SimRank 计算的代价为 $O(|V|^3)$,因此计算所有可能世界的代价为 $O(2^{|E|}|V|^3)$.在动态图的情况下,每给定一个查询点,求其 kNN,都需要重新对每个可能世界进行 SimRank 计算,这样代价非常昂贵.

4 优化算法

由于基本算法代价非常昂贵,因此有必要提出优化算法.

定义 5. 概率图对应的确定图.已知不确定图$\mathcal{G}=\langle V,E,P\rangle$和确定图 $G=\langle\bar{V},\bar{E}\rangle$,如果 $V=\bar{V}\wedge E=\bar{E}$,则称 G 为$\mathcal{G}$对应的确定图,这时,称$\mathcal{G}$为 G 对应的不确定图.

定义 6. 确定图 $G=\langle V,E\rangle$,节点 $a\in V$,$G'=\langle V',E'\rangle\subseteq G$ 且 $a\in V'$,r 为常数,则称 G'为 G 的以点 a 为中心、r 为半径的子图;简称为 G 的 r 半径子图.如果 G'满足以下条件:$\forall v\in V'$,在图 G 中总有 $shortestPath(a,v)\leqslant r$.这里 $shortestPath(a,v)\leqslant r$ 表示 v 到 a 的最短路径长度;称 r 为该子图的半径.

定义 7. 概率子图.已知不确定图$\mathcal{G}$和其对应的确定图 G,G'为 G 的以点 a 为中心、r 为半径的子图,$\mathcal{G}'$是$\mathcal{G}$的子图同时 G'为$\mathcal{G}'$对应的确定图,那么我们称$\mathcal{G}'$为$\mathcal{G}$的以点 a 为中心、r 为半径的概率子图简称子图.

定理 1. G 为确定图,G'为其上的以点 a 为中心、r为半径的子图,给定阈值α,C为SimRank公

式中的衰减因子. 当 $r=\left\lceil \log_{\sqrt{C}/2}\frac{\alpha(1-\sqrt{C}/2)C}{2}\right\rceil$ 时,$\forall b\in G-G'$,则 a 与 b 无关.

证明. 由文献[2]可知:

$$R_k(a,b)=\sum_{x=1}^{k}M_x(a,b), \tag{1}$$

这里 $M_k(a,b)$ 表示两个随机冲浪者分别从点 a,b 出发在第 k 步第 1 次相遇的概率:

$$Sim(a,b)=\lim_{k\to\infty}R_k(a,b). \tag{2}$$

而给定一个查询点 a 的 r 半径子图,如果给定一个不在子图中的点 b,即 $b\in G-G'$,那么 a 与 b 最早只能在第 $\left\lceil\frac{r}{2}\right\rceil$ 步上第 1 次相遇. 这时:

$$M_{\left\lceil\frac{t}{2}\right\rceil}(a,b)=C^{\left\lceil\frac{t}{2}\right\rceil}\sum_{t:a\sim b}P(t), \tag{3}$$

t 表示在第 $\left\lceil\frac{r}{2}\right\rceil$ 步上第 1 次相遇的路径,假设这样的路径个数是 k,而 $P(t)$ 表示相应路径的概率,而这样长度为 r 路径的 $P(t)$ 值最大的情况是除了始点与终点的度数为 k,路径上其他节点度数为 2,则有:

$$\sum_{t:a\sim b}P(t)\leqslant\frac{1}{k}\times\frac{1}{2^{r-1}}; \tag{4}$$

由式(3)(4)得:

$$M_{\left\lceil\frac{t}{2}\right\rceil}(a,b)\leqslant C^{\left\lceil\frac{t}{2}\right\rceil}\frac{1}{k\times 2^{r-1}}\leqslant$$

$$C^{\frac{r}{2}-1}\frac{1}{2^{r-1}}=\frac{2}{C}\times\left(\frac{\sqrt{C}}{2}\right)^{r}; \tag{5}$$

令 $y=\frac{\sqrt{C}}{2}$ 又因为 $M_{K+1}(a,b)<M_K(a,b)$,那么由式(1)(2)(3)(5)得:

$$Sim(a,b)=\lim_{k\to\infty}R_k(a,b)=$$

$$\lim_{x\to\infty}\sum_{x=1}^{k}M_x(a,b)\leqslant$$

$$\lim_{R\to\infty}\frac{2}{C}\sum_{m=r}^{R}(y)^m=\frac{2}{C}\frac{y^r}{1-y};$$

当 $\alpha>\frac{2}{C}\frac{y^r}{1-y}$ 时,可得 $r<\log_y\frac{\alpha(1-y)C}{2}$.

证毕.

定理 1 说明给定这样的子图 G' 与 a 相似的点都在子图中,因此点 a 的 kNN 点也在子图中. 例如当 $\alpha=0.01$,$C=0.5$ 时 $r=6$.

推论 1. 已知点 a 为中心、r 为半径的子图 G',给定阈值 α,C 为 SimRank 公式中的衰减因子. 当 $r=\left\lceil\log_{\sqrt{C}/2}\frac{\alpha(1-\sqrt{C}/2)C}{2}\right\rceil$ 时,$\forall b\in G-G'$,$\forall c\in G'$,子图外所有点 b 对值 $Sim(a,c)$ 的影响小于 α.

推论 1 说明给定子图 G',只需对子图计算所得到的 $Sim(a,c)$ 值与在原图中的真实值误差小于 α,只在子图中计算不会影响 kNN 查询.

由定义 5 至定义 7、定理 1、推论 1 可以得知,给定查询点 a 和阈值 α 可以得到对应的概率子图,该概率子图保证在所产生的每个可能世界里所计算出的点 a 和在概率子图中的其他点的 SimRank 值的误差不会超过 α,而且还保证了所有与点 a 的 SimRank 值大于 α 的点都在概率子图中.

定理 2. $\mathcal{G}$ 为 G 的以点 a 为中心、r 为半径的概率子图,α 为阈值. 对于任意 $b\in V(\mathcal{G})$,都有 $p_{a,b}(s)=p'_{a,b}(s)$. $p_{a,b}(s)$,$p'_{a,b}(s)$ 表示点 a 和点 b 的 SimRank 值取 s 时分别在不确定图 $\mathcal{G}$ 和 G' 中的概率.

本定理说明了不确定图中的概率不会影响优化算法. 对定理这里通过一个例子说明,图 1(a)是一个不确定图,虚线框中的子图是给定阈值 α 对应的查询点 a 的概率子图,图 1(b)是对应概率子图的一个可能世界,其概率是 $Pr_{(b)}=0.6\times0.5\times0.2=0.06$,假设在这个可能世界中 $S(a,b)=s_1$,那么在原不确定图中 a 和 b 的相似度值近似等于 s_1 的可能世界是图 1(c)(d). 而它们的存在概率分别为 $Pr_{(c)}=0.6\times0.5\times0.2\times0.3=0.018$,$Pr_{(d)}=0.6\times0.5\times0.2\times(1-0.3)=0.042$. 因此有 $Pr_{(b)}=Pr_{(c)}+Pr_{(d)}$. 符合定理 2.

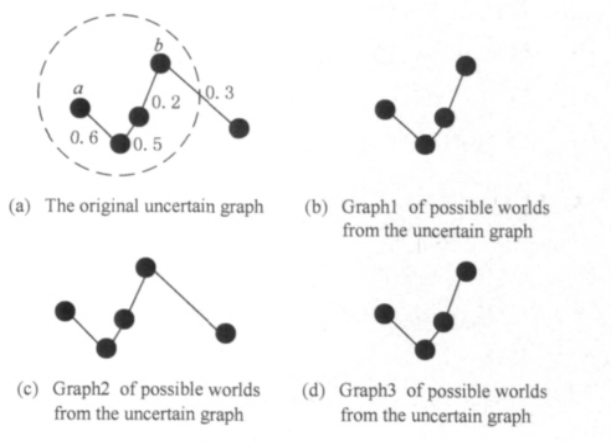


Fig. 1 The example of Theory 2.

图 1 定理 2 的例子

由本节的以上定义、定理得出给定阈值 α、查询点 a,$\mathcal{G}$为 G 的概率子图,对于任意节点 $v\in V(\mathcal{G})-V(\mathcal{G}')$,$v$ 都和 a 无关. 并且在概率子图$\mathcal{G}$中计算出的 S_M,S_J 和 S_{ER}的值和在$\mathcal{G}$中计算出的相应值的误差不超过 α.

由以上分析得出在不确定图$\mathcal{G}$上求解节点相似度的最近邻问题可转化为在其概率子图$\mathcal{G}'$上求解. 因此得到如下优化算法.

算法 2. 优化算法.

输入:不确定图$\mathcal{G}$、查询点 a,k、阈值 α;

输出:点 a 的 k 个最近邻节点.

① 执行 findSubGraph 算法在不确定图$\mathcal{G}$中获取以点 a 为中心、以 r 为半径的概率子图$\mathcal{G}'$.

② 在概率子图对应的所有可能世界中抽样出 m 个可能世界.

③ 对抽样出的每个可能世界中的所有点对计算出 SimRank 值,最后求出概率子图中的基于 SimRank 的概率度量值 M、返回集合 $T_K(a)$.

优化算法首先利用 findSubGraph 算法获取一个概率子图,这时问题转化为在这个概率子图上求 kNN,然后在概率子图上的每个可能世界上利用 1 节中介绍的 SimRank 的迭代计算的矩阵公式计算查询点和其他点的相似度值,最后求出 kNN.

在优化算法中采用了抽样技术来避免在概率子图的所有可能世界进行 SimRank 计算所带来的代价. 算法 findSubGraph 是采用图的广度优先搜索遍历算法的类似方式抽取子图的.

算法 3. findSubGraph 算法.

输入:图 G、查询点 a、常量 K 与 R;

输出:$SubG[]$.

```
/*位向量 visted 用来区分顶点是否被访问*/
bitset<N> visted;
InitQueue(Q); /*置空辅助队列*/
r=0; /*当前子图半径 r=0*/
subG.push_back(a); /*将顶点 a 插入有序数组中*/
visted.set(a); /*标识 a 已访问*/
EnQueue(Q,a); /*a 插入队列*/
EnQueue(Q,-1); /*用来统计子图半径*/
while((!Q.empty()) /*队列不为空*/
  DeQueue(Q,v);
  if(Q.empty()) break;endif
if(v!=-1)
    {p=g->adjL[v].next;
      while (p!=NULL)
      if(!visted[p->v]) /*顶点没访问过*/
        visted.set(p->v);
        EnQueue(Q,p->v);
        subG.push_back(p->v);
      endif
        p=p->next;
    endwhile
  }
else
  {/*子图半径加 1*/
    r++;EnQueue(Q,-1);
  if(r==R && !Q.empty() && subG.size()<k) /* 达到规定半径 R,但子图的节点度数小于 k 则 R 加 1*/
        R+=1;
      endif
        if(r>=R) break;endif
      }
endif
endwhile
```

findSubGraph 算法是给定图 G 的邻接表、查询点 a,按照图的广度优先搜索遍历算法的类似方式抽取出图 G 的以点 a 为中心、r 为半径的子图,与图的遍历算法不同的是本算法无需遍历图 G 中的所有点,只需遍历所要找的 G 的 r 半径子图中的所有节点. 从给定顶点 a 出发,记录 a 是子图中的顶点后,依次记录 a 的各个未曾访问过的邻接点为子图顶点,然后分别从这些邻接点出发依次记录它们的邻接点,直到子图的半径为 r. 然后根据子图的顶点信息构造对应的邻接矩阵.

在抽取子图过程中为了记录图 G 中的顶点是否为子图中的顶点,需设置一个有序数组 $SubG[]$,初值为 0,某个顶点 v_j 确定为子图中的顶点时,则把 j 插入这个有序数组中.

查询给定顶点的 kNN,当子图的节点数大于 k

且子图半径达到规定的半径时就停止查找,然而当子图的节点不足 k 且半径达到规定时,仍然需要继续扩展子图,直到子图的节点数超过 k 或者获得了整个连通分支为止.在实际抽取子图过程何时终止,必须考虑这种情况.

当得到子图 G' 的顶点数 m 和数组 $SubG[0..m-1]$,利用图 G 的邻接矩阵 W 快速得出子图的邻近矩阵 W':$W'[i,k]=W[SubG[i],SubG[k]]$,$\forall i,k$ 满足 $i,k\leqslant m$.

获取子图 $G'=\langle V',E'\rangle$ 时间复杂性为 $O(\mathrm{Max}((|V'|+|E'|),|V'|\mathrm{lb}|V'|)$.

优化算法可以解决 SimRank 只考虑节点间长度为偶数的路径,从而导致在有些情况下所得出的节点相似度为零的缺陷.通过对真实的论文的作者合作数据分析发现,作者之间的关系图形如图 2 所示,该图由 3 个连通分支组成,左侧大的连通分支表示大部分作者是连通的,右侧两个分支表示有些作者只和一个或两个作者合作过.通过 SimRank 计算后发现点 a,b 与其余所有节点相似度为 0,但通过图 1 发现点 a,b 肯定和他们所连通的点相似.给定查询点 a 和 b,优化算法发现 a,b 与其余所有节点相似度为 0 时,以 a,b 为中心的子图的节点是按从 a,b 出发距离由小到大且和 a 与 b 连通方式确定下来的.因此子图中其他点便是 a 和 b 的最相似的点,相似程度可通过距离确定,这样便解决了图 1 的问题.

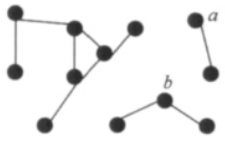


Fig. 2 A graph of coauthors' relation.

图 2 论文的作者合作图

下面两个引理保证了采用了抽样技术后结果的准确性.

引理 1. $\mathcal{G}=\langle V,E,P\rangle$ 是不确定图 $\mathcal{G}$ 的以点 a 为中心、r 为半径子图,α 为阈值.$\{G_i\}_P(1\leqslant i\leqslant m)$ 是基于 P 的 m 个可能世界的样本集合,给定节点对 (a,b),令 $G'\subseteq\{G_i\}_p$ 是 k 个图的集合且每个图中的 $Sim(a,b)\geqslant\alpha$,令 s_i 为对应的相似度值,当 $m>\frac{2\varepsilon^2}{(1-\alpha)^2}\ln(\frac{2}{\delta})$,有:$Pr(|\frac{1}{m}\sum_{G_i\in G'}s_i-S_{ER}(a,b)|\geqslant\varepsilon$.

证明由 Hoeffding Inequality 直接证出.

引理 2. 不确定图 $\mathcal{G}$ 的所有可能世界 N 中随机抽出 m 个样本:$g_1,\cdots,g_m$.$M=median(S_{g_1}(a,b),\cdots,S_{g_m}(a,b))$ 为样本中的中位数,S_M 为 a 和 b 在所有可能世界中相似度值期望的中位数,μ 为 $S_{ER}(a,b)$,C 为一个常数.当 $m\geqslant\frac{3C^2}{\varepsilon^2 m}\ln(\frac{2}{\delta})$ 时:

$$Pr(|M-S_M|\geqslant\varepsilon N)\leqslant\delta.$$

证明.存在一个常数 C,使得 $|M-S_M|=C|\frac{N}{m\mu}\sum_i^m S_{g_i}(a,b)-N|$,当 $|M-S_M|\geqslant\varepsilon N$,即 $C|\frac{N}{m\mu}\sum_{i=1}^m S_{g_i}(a,b)-N|\geqslant\varepsilon N$ 时,两边同时乘以 $\frac{\mu}{CN}$ 得:$|\frac{1}{m}\sum_{i=1}^m S_{gi}(a,b)-\mu|\geqslant\frac{\varepsilon}{C}\mu$.由 chernoff bound 定理得当 $m\geqslant\frac{3C^2}{\varepsilon^2\mu}\ln(\frac{2}{\delta})$ 时公式成立. 证毕.

文献[1]有类似的引理.

采用采样技术后整个优化算法的时间复杂性是包括抽取概率子图和在其上进行 kNN 查询,而抽取概率子图的时间代价只和子图的节点数和边数有关并且是线性的.这样优化算法的主要代价是在概率子图上计算 SimRank 值,假设样本数为 m,$V(\mathcal{G}')$,$E(\mathcal{G}')$ 为概率子图的顶点集合、边的集合,则这部分代价是 $O(m|V(G')|^3)$,m 的取值与子图的可能世界总数 $2^{|E(\mathcal{G}')|}$ 有关.所以优化算法的代价是由子图的边数和顶点数所决定的.

5 实验结果与分析

实验的运行环境为 Intel(R) Core(TM)2 Duo CPU E7500,2 GB 内存和 Windows 7 操作系统,文中算法采用 C++实现.

实验主要考察优化算法的效率和准确性,下面把优化算法称为子图算法.具体主要从 3 个方面衡量:子图节点数和原图节点数的比例、子图边数和原图边数的比例和准确率.因为 n 个节点的图上计算 SimRank 的代价是 $O(n^3)$,因此子图节点数所占比例越小性能越好.边数直接决定了可能世界的多少,虽然采用了样本技术,但样本总体数越少,在较少的采样数下能得到的结果越准确,采样数减少同时也减少了计算代价,所以边的比例可以衡量算法性能.而这两个比例的信息通过 findSubGraph 算法来获取,findSubGraph 算法的执行没有涉及到概率 P.

基于上述原因,这部分实验是在确定图上进行的,同时在此确定图和其子图上求给定查询点的 kNN,然后比较准确性,也可以反映出子图算法的准确性(无法在不确定图上进行准确的比较,而且子图算法和基本算法的最大不同是子图算法很大程度上减少了数据量,和概率无关,见定理 2).

子图算法本质上是剪枝技术,它得到的数据集只占原来的 10% 左右(见下面实验分析),其他的 SimRank 计算优化方法[14-15]也可以直接应用在子图算法得到的数据集上. 因此不失一般性,实验中计算 SimRank 是采用第 1 节介绍的矩阵迭代形式.

我们分别在两个真实的数据集上进行了实验. 这两个真实数据集来自斯坦福大学的网络分析平台[16],第 1 个数据集为 ca-GrQc,为作者的论文合作关系图,第 2 个数据集为 amazon0302,表示在亚马逊上当一个商品甲频繁地同商品乙一起被购买则甲与乙之间有边,其中 ca-GrQc 在其整个数据集进行实验,共有 5 242 个点,28 980 条边;而 amazon0302 上只是取其前 6 000 个点组成的图上进行实验,边数为 24 754. 为了简化实验我们取子图半径为 2,3 和 4,阈值分别为 0,0.001,0.01,0.1,$k=10$ 来分析子图算法效率和准确性.

图 3 至图 6 的每个数据分别是在两个数据集上随机抽取了 200 个查询点,然后对每个查询点进行子图算法后所得结果数值的平均值得到的,因此能比较准确反映出算法的性能. 例如错误率是这样计算出来的:200 个查询点中第 i 个点查询出的 k 个最近邻中有 m_i 个不是真正的最近邻,则错误率为 $\sum_{i=1}^{200} m_i/(200\times k)$. 由图 3 至图 6 可见,当半径、阈值越大,准确率越高,同时可以看出在保证准确率下,子图的点、边比例相对较小,因此子图算法是有效的.

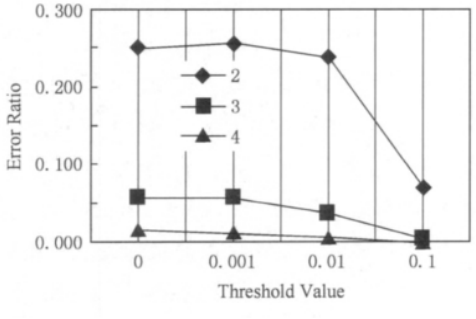


Fig. 3 The error ratio of kNN query on subgraph (ca-GrQc).

图 3 子图上求 kNN 的错误率(ca-GrQc)

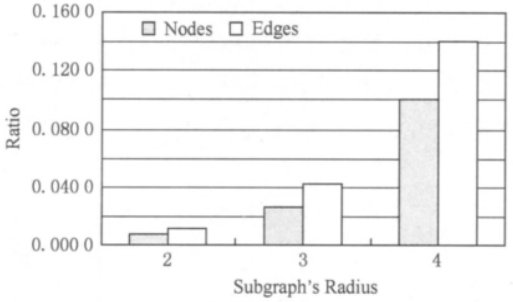


Fig. 4 The ratio of subgraph's nodes number and edges number to its original graphs' (ca-GrQc).

图 4 子图与原图的点、边比例(ca-GrQc)

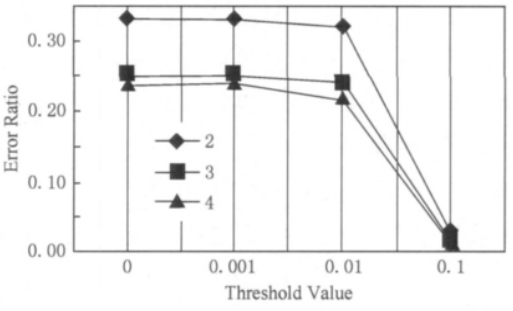


Fig. 5 The error ratio of kNN query on subgraph (amazon).

图 5 子图上求 kNN 的错误率(amazon)

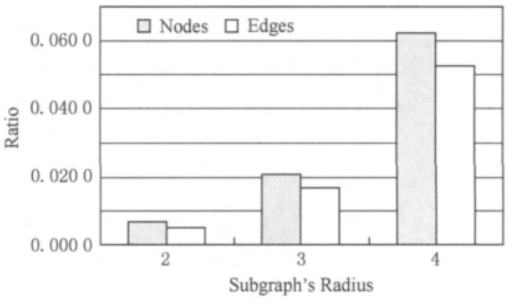


Fig. 6 The ratio of subgraph's nodes number and edges number to its original graphs' (amazon).

图 6 子图与原图的点、边比例(amazon)

第 2 部分实验是对 ca-GrQc 数据集上以正态分布 N(0.6,0.1)给每条边加概率得到一个不确定图. 然后在不确定图上以半径为 3 阈值为 0 基于 3 种 S_M, S_J, S_{ER} 度量进行子图算法求 kNN, k 取 5. 由图 3 可知取这样的半径和阈值是合理的,因为这时就可以保证准确率.

实验中随机生成 3 个查询点,在每个查询点的概率子图相应随机抽取 100 个可能世界,分别求出查询点的基于3种度量的kNN,查询时间如表1所示:

Table 1 Time of the Query in Uncertain Graph

表 1 不确定图上查询时间

Query Node(author number)	Query Time/s
1742	26.786
2669	0.078
859	58.646
Average time of query using Optimization algorithm	549
Time of query directly in original uncertain graph	134537

表 1 中的平均时间是查询随机抽取的 200 个查询点(在每个查询点对应的概率子图上同样取 100 个可能世界)基于 3 种度量的 kNN 的平均时间,表中最后一行的时间是指在整个不确定图上计算其中的 100 个可能世界的 SimRank 的时间,用这个时间代表基本算法时间(基本算法时间肯定大于这个值).由表中可以看出子图算法效率显然优于基本算法.

6 结 论

本文对非确定图上查询给定节点的 k 个最近邻问题,提出了基于 SimRank 度量的方法,虽然 SimRank 度量节点间的相似性准确性较高,但现有的 SimRank 在图动态演变的情况下查询最近邻代价较高,因此提出了一个优化算法,能够高效处理不确定图上的 kNN 查询处理.算法既可以作为计算 SimRank 的新方法同时又可以解决 SimRank 只考虑节点间长度为偶数的路径,从而导致在有些情况下所得出的节点相似度为零的缺陷.最后实验验证了算法的高效性和有效性.

参 考 文 献

[1] Michalis P, Francesco B, Aristides G. k-Nearest neighbors in uncertain graphs [J]. Proc of the VLDB Endowment, 2010, 3(1): 997-1008

[2] Hintsanen P, Toivonen H. Finding reliable subgraphs from large probabilistic graphs [J]. Data Mining and Knowledge Discovery, 2008, 17(1): 3-23

[3] Zou Zhaonian, Li Jianzhong, Gao Hong, et al. Frequent subgraph pattern mining on uncertain graph data [C] // Proc of the ACM Int Conf on Information and Knowledge Management. New York: ACM, 2009: 583-592

[4] Zou Zhaonian, Li Jianzhong, Gao Hong, et al. Finding top-k maximal cliques in an uncertain graph [C] //Proc of the Annual IEEE Int Conf on Data Engineering. Piscataway, NJ: IEEE, 2010: 649-652

[5] Zou Zhaonian, Gao Hong, Li Jianzhong. Discovering frequent subgraphs over uncertain graph databases under probabilistic semantics [C] //Proc of the ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2010: 633-642

[6] Zou Zhaonian, Li Jianzhong, Gao Hong, et al. Mining frequent subgraph patterns from uncertain graph data [J]. IEEE Trans on Knowledge and Data Engineering, 2010, 22(9): 1203-1218

[7] Gao Hong, Zhang Wei. Research status of the management of uncertain graph data [J]. Communications of the China Computer Federation, 2009, 5(4): 31-36 (in Chinese)
(高宏, 张炜. 不确定图数据管理研究现状[J]. 中国计算机学会通讯, 2009, 5(4): 31-36)

[8] Zhou Xun, Li Jianzhong, Shi Shengfei. Distributed aggregations for two queries over uncertain data [J]. Journey of Computer Research and Development, 2010, 47(5): 762-771(in Chinese)
(周逊, 李建中, 石胜飞. 不确定数据上两种查询的分布式聚集算法[J]. 计算机研究与发展, 2010, 47(5): 762-771)

[9] Liben-Nowell D, Kleinberg J. The link-prediction problem for social networks [C] //Proc of the ACM Int Conf on Information and Knowledge Management. New York: ACM, 2003: 556-559

[10] Jeh G, Widom J. SimRank: A measure of structural-context similarity [C] //Proc of the 8th ACM SIGKDD Int Conf on Knowledge discovery and Data Mining. New York: ACM, 2002: 538-543

[11] Loannis A, Hector G, Chang C C. Simrank++: Query rewriting through link analysis of the click graph [C] //Proc of the Int Conf on Very Large Data Bases. New York: ACM, 2008: 408-421

[12] Wang Jidong, Zeng Huajun, Chen Zheng, et al. ReCom: Reinforcement clustering of multi-type interrelated data objects [C] //Proc of the 26th Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2003: 274-281

[13] Yin Xiaoxin, Han Jiawei, Yu P S. LinkCLus: Efficient clustering via heterogeneous semantic links [C] //Proc of the Int Conf on Very Large Data Bases. New York: ACM, 2006: 427-438

[14] Li Pei, Liu Hongyan, Yu J X, et al. Fast single-pair SimRank computation [C] //Proc of the SIAM Int Conf on Data Mining. Philadelphia: SIAM, 2010: 571-582

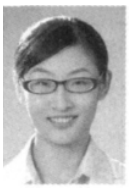
[15] Li Cuiping, Han Jiawei, He Guoming, et al. Fast computation of SimRank for static and dynamic information networks [C] //Proc of the 13th Int Conf on Extending Database Technology. New York: ACM, 2010: 465-476

[16] Standford Large Network Dataset Collection [EB/OL]. [2011-05-10]. http://snap.stanford.edu/data/

Zhang Yinglong, born in 1979. He is a PhD candidate at Renmin University of China. His research interests include data mining, etc.

Li Cuiping, born in 1971. Received her PhD degree from Chinese Academy of Sciences in 2003. Associate professor and PhD supervisor at Renmin University of China. Her research interests include database, data mining, information network analysis and data stream management, etc.

Chen Hong, born in 1965. Received her PhD degree from Chinese Academy of Sciences in 2000. Professor and PhD supervisor at Renmin University of China. Her research interests include database, data mining, data stream analysis and management, sensor network data management, etc.



Du Lingxia, born in 1986. MSc candidate at Renmin University of China. Her research interests include data mining, etc.