软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2018,29(10):3150−3163 [doi: 10.13328/j.cnki.jos.005286]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

一种不确定图数据库上的相似性连接方法*

缪丰羽[1], 王宏志[2]

[1](宁德师范学院 信息与机电工程学院,福建 宁德 352100)

[2](哈尔滨工业大学 计算机学院,黑龙江 哈尔滨 150001)

通讯作者: 缪丰羽, E-mail: feathen1983@163.com

摘 要: 在确定图上进行的相似性连接已有许多研究成果.然而,在实际应用中会有许多因素使得图结构数据变得不确定.研究了不确定图数据库上的相似性连接问题.采用联合概率分布表示法来描述图中边的不确定性,结合一种新的图的相似性度量方法,给出了不确定图数据库上的相似性连接的形式化定义,并设计了一组过滤策略来减少连接过程中候选图对的数量.大量的实验数据表明,所提出的方法具有较好的可行性和准确性.

关键词: 不确定图;联合概率分布;相似性连接;过滤策略

中图法分类号: TP311

中文引用格式: 缪丰羽,王宏志.一种不确定图数据库上的相似性连接方法.软件学报,2018,29(10):3150–3163. http://www.jos.org.cn/1000-9825/5286.htm

英文引用格式: Miao FY, Wang HZ. Method for similarity join on uncertain graph database. Ruan Jian Xue Bao/Journal of Software, 2018,29(10):3150−3163 (in Chinese). http://www.jos.org.cn/1000-9825/5286.htm

Method for Similarity Join on Uncertain Graph Database

MIAO Feng-Yu[1], WANG Hong-Zhi[2]

[1](College of Information & Mechanical and Electrical Engineering, Ningde Normal University, Ningde 352100, China)

[2](School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

Abstract: Many studies have been conducted on similarity join over certain (deterministic) graphs. However, in reality, graphs are often uncertain due to various factors. This paper studies similarity join on uncertain graph databases. The study employs the joint probability distribution to describe the uncertainty of edges in the graph, combines a new measure to evaluate graph similarity, and gives the formal definition of the similarity join on uncertain graph database. The paper also designs a group of filtering strategies to reduce the candidate pairs in the similarity join. A large number of experimental data show that, the method proposed in the paper is feasible and accurate.

Key words: uncertain graph; joint probability distribution; similarity join; filtering strategy

图模型被广泛应用于各种复杂的数据表示,例如生物信息、社会网络、本体网络、XML 数据和 RDF 数据等.由于现代科学的研究方法和测量技术普遍存在着误差和噪声,图数据中经常会引入不确定信息,由此产生了不确定社会网络、不确定本体网络、不确定道路交通网络、不确定 XML 数据和不确定 RDF 数据等数据模型[1].目前,大部分的不确定图研究采用的是概率图模型[2,3].在这种模型中,图的每条边被赋予一个概率值来定量描述这条边存在于图中的可能性,并且边的概率值互相独立.然而,在大部分的实际情况中,图中边存在的可

* 基金项目: 国家科技支撑计划(2015BAH10F00); 国家自然科学基金(61472099, 61133002); 福建省自然科学基金(2018J01555); 福建省教育厅中青年项目(JAT170653); 宁德师范学院校级青年专项基金(2014Q51)

Foundation item: National Key Technology Support Program of China (2015BAH10F00); National Natural Science Foundation of China (61472099, 61133002); Natural Science Foundation of Fujian Province, China (2018J01555); Fujian Provincial Department of Education Youth Project (JAT170653); University Youth Foud of Ningde Normal University (2014Q51)

收稿时间: 2016-03-11; 修改时间: 2016-06-03, 2016-08-11; 采用时间: 2017-03-17; jos 在线出版时间: 2017-07-12

CNKI 网络优先出版: 2017-07-12 15:33:36, http://kns.cnki.net/kcms/detail/11.2560.TP.20170712.1533.006.html

能性是相互关联的.例如,在通信网络中用边的概率来表示一个连接的可靠性,这些网络的路由路径之间显然是相互关联的;又如,道路交通网络中用边的概率来表示某路段的交通拥堵程度,一个繁忙的交通路段也经常会造成附近通路的阻塞.因此,我们需要选择一种新的模型来描述这种相互关联的不确定性.概率论中的联合概率分布可以反映若干个相互关联的随机变量的概率分布情况,因此,本文选择了联合概率分布表示法[4]来描述不确定图中边与边之间的相互关联.

在图数据管理中,相似性连接问题是目前研究的热点之一.图的相似性连接指的是从一个图数据库中找出所有相似的图对,这些图对之间的差异小于给定的阈值.图之间的相似性通常用图的编辑距离[5]来表示.图的编辑距离是指将一个图转化成另一个图需要进行的基本操作的数目.可执行的基本操作包括[6]:(1) 插入一个顶点;(2) 删除一个顶点及其所有边;(3) 修改顶点的标签;(4) 在两个不相连的顶点之间插入一条边;(5) 删除一条边;(6) 修改边的标签.图之间的编辑距离越小,则相似性越大.然而,计算图的编辑距离是 NP-完全问题.因此,人们提出了各种近似算法.但这些算法很难集高时效、高质量和简单于一身.图的编辑距离主要强调图之间的整体差异,然而,在很多具体的应用中,更侧重于以顶点为中心的局部相似性.例如,在社会网络中,插入或者删除一个朋友,主要影响围绕该顶点的子结构;又如,在计算机视觉中,图模型之间的匹配可以通过围绕相应顶点的子结构之间的相似性匹配来计算.文献[7]提出了一种新的相似性度量方法,通过计算两个图之间的匹配率来进行相似性查询.这种新的度量方法适合于那些以顶点为中心的子结构相似性应用,具有广泛的适用性.

由于各种原因造成的图数据的不确定性导致人们在实际应用中需要处理的大部分图都是不确定图,而相似性连接又是图数据管理中最重要的操作之一,因此,针对不确定图进行的相似性连接成为了图数据管理中亟待解决的问题.例如,在社会网络中,由于更新不及时,可能出现相同的社区具有若干不同的关系或标签,若采用不确定图进行建模,则相似性连接结果可能会包含该社区对.这相比于确定图中的相似性连接更具有容错性,也更接近于人的分析判断方式.但到目前为止,大部分相似性连接问题的研究都是基于确定图进行的.由于不确定图中边的存在具有不确定性,一个不确定图可能对应着若干个确定图,且每个确定图都有相应的存在概率,因此,基于确定图的相似性连接算法不能直接应用于不确定图.本文针对不确定图的特征,选择了一种联合概率分布表示法来表达图中的不确定性,并采用文献[7]提出的匹配率计算方法来描述图之间的相似性,从而计算一个不确定图数据库中所有满足给定阈值的相似图对.实验结果表明,本文提出的方法具有较好的可行性和较高的准确性.

本文的主要贡献包括:

(1) 将图的相似性连接问题从确定图的范畴扩展到了不确定图.据我们所知,目前在这个课题上的成果还很少.

(2) 给出了不确定图数据库上的相似性连接的形式化定义,并从时间复杂度上进行了详细的分析.

(3) 设计了一组过滤算法,大大减少了连接过程中的候选图对的数量.

(4) 给出了不确定图数据库上的相似性连接算法以及相应的时间复杂度分析.

(5) 通过实验,其结果表明了本文所提出的相似性连接方法的可行性和有效性.

本文第 1 节介绍基于联合概率分布表示法的不确定图模型.第 2 节给出不确定图数据库上的相似性连接的形式化定义.第 3 节介绍本文提出的 4 个过滤策略.第 4 节给出相似性连接的算法以及时间复杂性分析.第 5 节介绍在本文提出的算法上进行的一系列实验及相应的分析.第 6 节为总结和展望.

1 基于联合概率分布的不确定图模型

本节介绍了基于联合概率分布表示法的不确定图模型.首先介绍了确定图的定义,并基于该定义给出了本文所讨论的不确定图的形式化定义,最后给出了可能世界图的定义.基于可能世界图,不确定图和确定图之间可以互相转化.以上 3 个定义都参考于文献[4].

定义 1(确定图). 一个确定图 $g^c=(V,E,\Sigma,L)$,其中,V 是顶点集,E 是边集,Σ是标签集,$L:V\cup E\rightarrow\Sigma$是一个函数,该函数将标签映射到顶点和边上.如果 g^c 中的某些边射入同一个顶点或者组成一个三角形,我们就把它们称为邻

边,记为 *ne*.

定义 2(不确定图). 一个不确定图 $g=(g^c,X_E)$,其中,g^c 为一个确定图,X_E 是一个由 E 索引的二进制随机变量集.X_E 中的元素 x_e 取值为 0 或者 1,用来表示边 e 存在的概率.图中的每个邻边集都有一张对应的联合概率分布表(JPT),表中给出了该邻边集中邻边的所有随机变量分布情况以及相应的概率值 *Pr*.

例如:图 1 所示的不确定图有 3 个邻边集,分别是($e1$,$e2$,$e3$)、($e3$,$e4$,$e5$)和($e4$,$e5$,$e6$).对应的联合概率分布表为图 2 中的 JPT1、JPT2、JPT3.JPT1 中的第 2 行数据 $Pr(e1=1,e2=1,e3=0)=0.2$ 表示"$e1$、$e2$ 边存在,$e3$ 边不存在"的概率为 0.2.

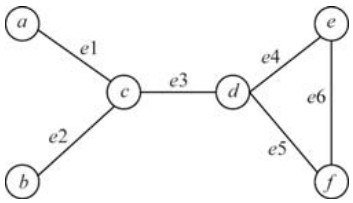

Fig.1 Uncertain graph

图 1 不确定图

e1	e2	e3	Pr
1	1	1	0.2
1	1	0	0.2
1	0	1	0.1
1	0	0	0.1
0	1	1	0.1
0	1	0	0.1
0	0	1	0.1
0	0	0	0.1

JPT1

e3	e4	e5	Pr
1	1	1	0.3
1	1	0	0.4
1	0	1	0.05
1	0	0	0.15
0	1	1	0.03
0	1	0	0.02
0	0	1	0.04
0	0	0	0.01

JPT2

e4	e5	e6	Pr
1	1	1	0.25
1	1	0	0.15
1	0	1	0.05
1	0	0	0.15
0	1	1	0.08
0	1	0	0.2
0	0	1	0.04
0	0	0	0.08

JPT3

Fig.2 JPTs

图 2 JPTs

我们讨论的不确定图含有不确定边,但顶点确定,这符合大部分实际应用的情况.显然,在一个大图中,对整个顶点集 V 建立联合分布表是不切实际的.在本文中我们只针对邻边建立联合分布表,因为在实际情况中,邻边的相互影响最为明显.

定义 3(可能世界图). 一个可能世界图 $g'=(V',E',\Sigma',L')$是一个确定图,它是一个不确定图 $g=((V,E,\Sigma,L),X_E)$的实例,其中,$V'=V,E'\subseteq E,\Sigma'\subseteq\Sigma$.

每个可能世界图都有一个相应的概率,其值为每个 JPT 中相应的元组的 *Pr* 值的乘积.一个不确定图的所有可能世界图的概率的和小于等于 1.

例如,将 JPT1 中的元组($e1=1,e2=1,e3=1$)与 JPT2 中的元组($e3=1,e4=1,e5=1$)以及 JPT3 中的元组($e4=1,e5=1,e6=1$)进行联接,会得到图 1 所示的一个可能世界图,它表示图 1 中所有的顶点和边都存在.该可能世界图的概率值为 $0.2\times0.3\times0.25=0.015$.图 1 所示可能世界图的个数为 3 个联合分布表进行联接得到的表的元组数.

2 不确定图数据库上的相似性连接

受文献[8]的启发,下面我们给出不确定图数据库上的相似性连接的形式化定义.

定义 4(不确定图数据库上的相似性连接,USJ). 给定一个不确定图数据库 $G=\{g_1,g_2,\ldots,g_N\}$,一个相似性阈值$\gamma\in(0,1]$,一个概率阈值$\alpha\in(0,1]$,不确定图数据库上的相似性连接指的是从 G 中找出所有 *Pr* 值大于等于α的图对(g_i,g_j).

$$USJ=\{(g_i,g_j)\mid g_i\in G \text{ and } g_j\in G \text{ and } Pr\{Sim(g_i,g_j)\geqslant\gamma\}\geqslant\alpha\} \tag{1}$$

其中,

$$Pr\{Sim(g_i,g_j)\geqslant\gamma\}=\sum_{\forall g'\in g_i}\sum_{\forall g''\in g_j}p(g')\times p(g'')\times\chi(Sim(g',g'')\geqslant\gamma) \tag{2}$$

其中,g'为 g_i 的可能世界图,g''为 g_j 的可能世界图,$p(g')$为 g'的概率,$p(g'')$为 g''的概率.

$$\chi(Sim(g',g'')\geqslant\gamma)=\begin{cases}1, & \text{if } Sim(g',g'')\geqslant\gamma\\ 0, & \text{otherwise}\end{cases} \tag{3}$$

其中,$Sim(g',g'')$为g'和g''的相似度.图的相似性度量有很多种,这里我们采用的是文献[7]提出的匹配率度量方法.计算过程详见本文第 4.3 节.

例:给定一个不确定图数据库 G,G 上的相似性连接需要返回所有满足定义 4 的图对.给定相似性阈值γ=0.8,概率阈值α=0.5.设(g_1,g_2)为 G 中的一个图对,其中的不确定图 g_1、g_2 及其相应的联合概率分布表 JPT 分别如图 3~图 6 所示.

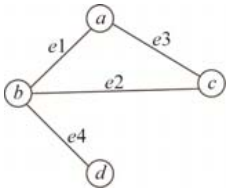

Fig.3 g_1

图 3 g_1

e1	e2	e3	Pr
1	1	1	0.8
1	1	0	0
1	0	1	0.2
1	0	0	0
0	1	1	0
0	1	0	0
0	0	1	0
0	0	0	0

JPT1

e1	e2	e3	Pr
1	1	1	0.7
1	1	0	0
1	0	1	0.3
1	0	0	0
0	1	1	0
0	1	0	0
0	0	1	0
0	0	0	0

JPT2

Fig.4 JPTs of g_1

图 4 g_1 的 JPTs

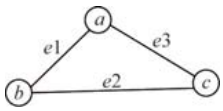

Fig.5 g_2

图 5 g_2

e1	e2	e3	Pr
1	1	1	0.6
1	1	0	0
1	0	1	0.4
1	0	0	0
0	1	1	0
0	1	0	0
0	0	1	0
0	0	0	0

JPT

Fig.6 JPT of g_2

图 6 g_2 的 JPT

我们首先根据 g_1、g_2 的 JPT 计算出其所有的可能世界图和相应的概率,如图 7、图 8 所示.

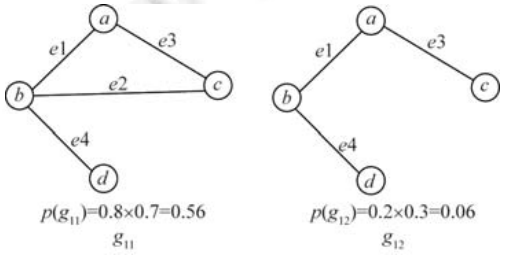

Fig.7 The PWGs and their probabilities of g_1

图 7 g_1 的可能世界图和相应的概率

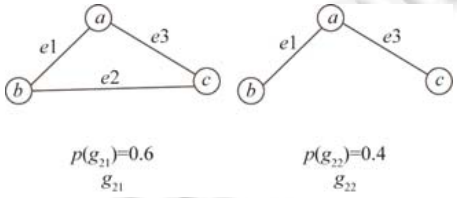

Fig.8 The PWGs and their probabilities of g_2

图 8 g_2 的可能世界图和相应的概率

然后计算 g_1、g_2 的所有可能世界图对的相似度.计算结果如下.

$Sim(g_{11},g_{21})$=7/8=0.875、$Sim(g_{11},g_{22})$=7/8=0.875、$Sim(g_{12},g_{21})$=7/12≈0.583、$Sim(g_{12},g_{22})$=5/6≈0.833.只有 3 个图对(g_{11},g_{21})、(g_{11},g_{22})和(g_{12},g_{22})的相似度大于给定的相似性阈值 0.8.

根据公式(2)和公式(3),

$$\begin{aligned}Pr\{Sim(g_1,g_2)\geqslant\gamma\}&=p(g_{11})\times p(g_{21})+p(g_{11})\times p(g_{22})+p(g_{12})\times p(g_{22})\\&=0.56\times0.6+0.56\times0.4+0.06\times0.4\\&=0.584\\&>\alpha.\end{aligned}$$

因此,该不确定图对在给定的阈值条件下相似.

根据定义 4,我们给出不确定图数据库中相似性连接的时间复杂性分析:对于不确定图数据库 $G=\{g_1,g_2,\ldots,g_N\}$,其中的图对数目为 N^2.根据公式(2),对于任意一个图对(g_i,g_j),其 Pr 值的计算需要统计 g_i 和 g_j 中所有的可能世界图对的概率乘积之和.设不确定图 g 中的邻边集个数为 t,则对应的 JPT 数为 t,假定邻边集中的边数为 r,则 JPT 中的元组数为 2^r,连接得到的可能世界图的数目为 2^{tr}.因此,g_i 和 g_j 中所有的可能世界图对为 2^{2tr} 个.公式(3)中的 $Sim(g',g'')$为相似性连接中的基本计算,设可能世界图中顶点和边的数目分别是 n 和 m,根据文献[7]的分析,$Sim(g',g'')$的时间复杂度为 $O(n\log n+m)$.综上所述,不确定图数据库上的相似性连接的时间复杂度在最坏情况下为 $O(N^2\times 2^{2tr}\times(n\log n+m))$.

3 过滤策略

根据定义进行不确定图数据库上的相似性连接,其时间复杂度为指数级,尤其是随着不确定图中邻边集数量的变化,可能世界图的数量也会产生较大的变化.而实际情况中,真正能够满足相似性阈值的图对是很少的.因此,我们设计了一组过滤策略,在算法运行的各个阶段进行过滤,以减少最后进行匹配率计算的可能世界图对,提高算法的运行效率.

3.1 单图上界过滤策略

对于给定的不确定图数据库 G,如果可以在算法运行初期,先筛选掉一部分不可能找到相似图的不确定图,就可以减少大量的候选图对,从而大大减轻后期的无效计算.

通过对第 3 节相关公式的分析和推导,我们得出了不确定图数据库中任意图上的一个定理.

定理 1. 设 g_i 为不确定图数据库 G 中的任意不确定图,g'为 g_i 的可能世界图,$p(g')$为 g'的概率.若 g_i 满足条件: $\sum_{\forall g'\in g_i} p(g')<\alpha$, 则可以判断 g_i 在 G 中没有满足条件的相似图,可以被剪枝.

证明:根据公式(2),我们有:

$$\begin{aligned}Pr\{Sim(g_i,g_j)\geqslant\gamma\}&=\sum_{\forall g'\in g_i}\sum_{\forall g''\in g_j}p(g')\times p(g'')\times\chi(Sim(g',g'')\geqslant\gamma)\\&\leqslant\sum_{\forall g'\in g_i}\sum_{\forall g''\in g_j}p(g')\times p(g'')\\&=\sum_{\forall g'\in g_i}p(g')\times\sum_{\forall g''\in g_j}p(g'').\end{aligned}$$

由可能世界图的性质可知, $\sum_{\forall g''\in g_j}p(g'')\leqslant 1$, 因此,

$$Pr\{Sim(g_i,g_j)\geqslant\gamma\}=\sum_{\forall g'\in g_i}p(g')\times\sum_{\forall g''\in g_j}p(g'')\leqslant\sum_{\forall g'\in g_i}p(g').$$

定理得证. □

根据上述定义得到单图上界过滤策略,若不确定图 g_i 的所有可能世界图的概率之和小于α,则将其进行过滤.设不确定图数据库 G 中的不确定图有 m 个,如果过滤掉 1 个没有希望的不确定图,则可以减少 $m-1$ 个候选图对,如果过滤掉 2 个没有希望的不确定图,则可以减少$(m-1)+(m-2)$个候选图对…,如果过滤掉 k 个没有希望的不确定图,则可以减少 $\sum_{i=1}^{k}m-i$ 个候选图对.

3.2 图对上界过滤策略

根据单图上界过滤策略过滤掉部分不确定图之后,剩余的不确定图将组成候选图对.如果在这些候选图对中可以进一步过滤掉一些不可能相似的图对,也会减少大量的后期计算.

在分析第 3 节相关公式的过程中,我们进一步得出了有关候选图对的一个定理.

定理 2. 设(g_i,g_j)为不确定图数据库 G 中的任意一个图对,g'为 g_i 的可能世界图,g''为 g_j 的可能世界图,$p(g')$

为 g'的概率,$p(g'')$为 g''的概率.若(g_i,g_j)满足条件: $\sum_{\forall g'\in g_i} p(g')\times \sum_{\forall g''\in g_j} p(g'')<\alpha$, 则可以判断图对$(g_i,g_j)$不满足相似性条件,可以被剪枝.

证明:从定理 1 的证明过程可知:

$$Pr\{Sim(g_i,g_j)\geqslant\gamma\}=\sum_{\forall g'\in g_i} p(g')\times \sum_{\forall g''\in g_j} p(g''),$$

结合公式(2),定理 2 得证. □

根据上述定理,得到图对上界过滤策略,若图对(g_i,g_j)的所有可能世界图对的概率乘积之和小于α,则将其进行过滤.

图对上界过滤策略通过计算所有可能世界图对的概率乘积之和来估计一个图对的相似性上界,避免了对所有可能世界图对的相似性进行的盲目计算.过滤图对的数量取决于具体的不确定图数据库.实验结果表明,在一般情况下,该步骤的过滤策略可以减少大量的相似性计算.

3.3 顶点权值过滤策略

在根据定义 4 进行相似性连接时,需要计算每个可能世界图对的相似度 $Sim(g',g'')$.本文采用了文献[7]中提出的相似性度量方法,这里先简单介绍一下该方法的计算过程.

首先给出一些定义和说明:在一个确定图中,$Con(v_i)$表示图中顶点 v_i 的度;顶点 v_i 的权值 $w(v_i)$为顶点 v_i 的度 $Con(v_i)$除以图中所有顶点的度之和 $Con(V)$,其中,V 为图中的顶点集;顶点 v_i 的连通集 $CS(v_i)$为顶点 v_i 所有邻接顶点组成的集合;顶点映射指的是一个图中的一个顶点 u 能够找到另一个图中具有相同标签的顶点 v 与之匹配,记为$\rho(u)=v$;两个顶点 u 和 v 连通集的相似性定义为 $COM(u\cap v)$,为 $CS(u)$和 $CS(v)$中映射顶点对的最大数目.

两个确定图 g'和 g''的相似度的计算步骤如下(假设 g'的顶点集为 V_1,g''的顶点集为 V_2,且 g'为两个图中顶点数多的那个图).

1) 计算 V_1 中所有顶点的权值和连通集,计算 V_2 中所有顶点的连通集.

2) 对于 V_1 中的每个顶点,找出其在 V_2 中的所有映射,我们将 V_1 中的顶点分成 Y 和 N 两个部分,Y 中的顶点可以在 V_2 中找到映射,N 中的顶点不行.

3) 对于 Y 中的每个顶点 v,在 V_2 中有顶点 u 与之映射,分为以下两种情况.

(1) 顶点 v 的标签在 Y 中是唯一的,且 V_2 当中 u 的标签在 V_2 中也是唯一的.这种情况下,顶点 u 和 v 可以匹配,我们可以直接计算 $COM(u\cap v)$.

(2) 顶点 v 的标签在 Y 中不是唯一的,或者 u 的标签在 V_2 中不是唯一的.在这种情况下,令 $V=\{v,v_1,v_2,\ldots,v_m\}$为 Y 中与 v 的标签相同的所有顶点的集合,$U=\{u,u_1,u_2,\ldots,u_n\}$为 V_2 中与 u 的标签相同的所有顶点的集合,再将 U 和 V 中的顶点按照顶点的度数降序排列.接下来,会有两种可能的情况.

① 若 $m\leqslant n$,对于 V 中的每个顶点 v_i,按照排好的顺序,寻找一个 U 中未匹配的顶点 u_j,使得 $COM(u_j\cap v_i)/|CS(u_j)|$的值最大.注:$U$ 中的每个顶点只能被匹配 1 次.

② 否则,对于 U 中的每个顶点 u_j,按照排好的顺序,寻找一个 V 中未匹配的顶点 v_i,使得 $COM(u_j\cap v_i)/|CS(u_j)|$的值最大.注:$V$ 中的每个顶点只能被匹配 1 次.

最后得到 g'和 g''的相似度为

$$Sim(g',g'')=\sum_{v\in Y} W(v)\times Com(v\cap u)/|CS(u)| \tag{4}$$

文献[7]在计算相似度时使用了一种顶点相似性上界过滤策略.该策略的原理是:从计算相似度的相关定义和说明可以看出,公式(4)中的 $COM(u\cap v)/|CS(u)|$的值不会超过 1.当这个值为 1 时,只有顶点的权值可以影响最终的结果,由此可以得到相似度的上界.定义顶点相似性为 Y 中顶点的权值之和.顶点相似性也就是图对相似度的上界.如果该上界比连接阈值要小,则可以断定该图对不相似.

将该策略进行简单的修改,可以得到不确定图数据库上的相似性连接所使用的顶点权值过滤策略.

顶点权值过滤策略:设可能世界图对(g',g'')在计算相似度的过程中产生的 Y 集合中的顶点的权值之和为

W_Y,若 $W_Y<\gamma$,则 $\chi(Sim(g',g'')\geqslant\gamma)=0$, 即$(g',g'')$可以从计算 Pr 值的所有可能世界图对中过滤掉.

这个过滤策略可以在很大程度上减少计算可能世界图对相似度时的计算量.过滤的可能世界图对的数目由具体的数据决定.但从我们的实验可以看出,这个环节的过滤取得了明显的效果.

3.4 不匹配顶点过滤策略

在计算相似度的过程中,若步骤 3 中的 V 或者 U 中还剩余一些不能匹配的顶点,此时还可以进行二次过滤.因为这些不匹配顶点的权值对相似度的最终结果没有贡献,因此也可以将其从 W_Y 中扣除,以求得更紧的可能世界图对的相似性上界.

不匹配顶点过滤策略:设可能世界图对(g',g'')在计算相似度过程中剩余的不匹配顶点的权值之和为 W_{left},若 $W_Y-W_{left}<\gamma$,则 $\chi(Sim(g',g'')\geqslant\gamma)=0$, 即$(g',g'')$可以从计算 Pr 值的所有可能世界图对中过滤掉.

该过滤策略可以实现对于相似度计算的进一步优化.通常,在计算相似度的算法中结合顶点权值过滤策略一起使用,可以尽可能地减少可能世界图对相似度的计算量.

4 算法及时间复杂性分析

综合上述剪枝策略,本节我们给出不确定图数据库上的相似性连接方法的算法描述及时间复杂性分析.首先介绍算法框架,然后介绍几个主要模块中的算法,最后综合几个模块,介绍整个算法.

4.1 算法框架

根据本文第 2 节给出的定义,不确定图数据库上的相似性连接返回的是不确定图数据库中满足相应阈值条件的所有图对.因此,连接算法主要是通过计算每个图对的 Pr 值并与概率阈值α相比较来判断连接结果.在计算不确定图对 Pr 值的过程中,需要先计算不确定图对所有的可能世界图及其相应的概率,因此必须先对输入的图数据构造出联合概率分布表 JPT.结合第 3 节提出的一组过滤策略,可以将本文提出的不确定图数据库上的相似性连接算法分为以下 6 个基本步骤.

(1) 根据输入的图数据以及每条边上的概率值,构建不确定图数据库.其中,每个不确定图根据邻边集的数目构造相应数量的 JPT.细节将在第 4.2 节中加以讨论.

(2) 将每个不确定图的 JPT 进行联接,计算出其所有的可能世界图以及相应的概率值.

(3) 将每个不确定图的所有可能世界图的概率值相加,若相加之和小于概率阈值α,则将其进行过滤,否则,将该图加入第 1 阶段的候选集 $C1$ 中.细节将在第 4.3 节中加以讨论.

(4) 检查第 1 阶段候选集 $C1$ 中的所有图对,若图对(g_i,g_j)满足定理 2 中的条件,则将其进行过滤,否则,将图对(g_i,g_j)加入第 2 阶段的候选集 $C2$ 中.细节将在第 4.4 节中加以讨论.

(5) 根据公式(2)计算 $C2$ 中的所有图对的 Pr 值.在计算过程中,根据顶点权值过滤策略和不匹配顶点过滤策略过滤掉一部分没有贡献的可能世界图对,细节将在第 4.5 节中加以讨论.根据剩余的可能世界图对的 Sim 值计算每个图对的 Pr 值.

(6) 将 Pr 值大于α的图对作为相似性连接的结果返回.

4.2 JPT生成算法

算法 1 用来生成每个不确定图的联合概率分布表 JPT.算法的输入为每个不确定图的顶点集和边集以及每条边上的存在概率,输出为不确定图对应的一组 JPT.

在后续的实验中,我们将采用真实的有机物 PPI 网络作为测试数据.一个 PPI(protein-protein interaction)网络是一个不确定图,其中的顶点代表蛋白质,边代表蛋白质间的相互作用,边上的存在概率 P 代表了相互作用的强弱.文献[9]证实了 PPI 的一个重要特点:一组相邻的相互作用受其中最强的作用支配,即 PPI 网络中每个邻边集的概率由该邻边中最大的概率值确定.因此,对于不确定图中的每个邻边集 ne,我们设置它们的概率为该邻边集中最大的概率值.例如,在图 1 所示的不确定图中,若有 $P(e1)=0.3$,$P(e2)=0.7$,$P(e3)=0.5$,则可以构造出相应的联合概率分布表 JPT1.其中,$Pr(e1=1,e2=1,e3=1)=\text{Max}(0.3,0.7,0.5)=0.7$,$Pr(e1=1,e2=1,e3=0)=\text{Max}(0.3,0.7)=0.7$,$Pr$

($e1$=1,$e2$=0,$e3$=1)=Max(0.3,0.5)=0.5…

算法的第 1 行将读入的顶点集、边集以及边上的概率值保存在邻接表中.算法的第 2 行~第 18 行对不确定图中的每个顶点分别进行判断:第 4 行~第 9 行处理第 1 种情况,即有多条边入射顶点 v.这种情况需要构建一个邻边集,设 v 的邻边数为 r,则 JPT 中的行数为 2^r,JPT 的列数为 r+1.接下来为每行填入相应的 Pr 值.第 10 行~第 18 行处理第 2 种情况,即 v 存在于一个三角形中.这种情况也需要构建一个邻边集,且该邻边集对应的 JPT 中的行数为 8,列数为 4.每个顶点的 *flag* 值用来标记该顶点是否已经构建过 JPT,以避免重复处理.第 19 行返回所有构造完的 JPT.

算法 1. Generate JPTs.

输入:V: nodes of an uncertain graph;E: edges of an uncertain graph;P: existence probability of the edges;

输出:JPTs.

1. input (V,E,P), and stored in an adjacency list A //读入不确定图的顶点集、边集和每条边上的概率值,存储于邻接表中
2. **for** each node v_i in the V **do**
3. v_i.*flag*=0;
4. **if** (|edges of v_i|≥3 and v_i.*flag*=0) **then** //当 v_i 所连接的边数大于等于 3 且 v_i 不属于任何已创建 JPT 中时,创建 v_i 所在的邻边集对应的 JPT
5. create an empty JPT for v_i (the number of rows are 2^r, the number of columns are r+1);
6. **for** each row in the JPT **do** //JPT 中每一行的 Pr 值为该行存在的边中的最大概率值
7. write the existence of edges in the JPT (‘0’ for not exist and ‘1’ for exist);
8. write the value of Pr (the max of probability of exist edges) in the JPT;
9. v_i.*flag*=1; *JPTs*=*JPTs*∪*JPT*;
10. **else if** (|edges of v_i|=2 and v_i is in a triangle and *flag*=0) **then** //当 v_i 所连接的边数等于 2、v_i 存在于一个三角形中且 v_i 不属于任何已创建 JPT 中时,创建 v_i 所在的邻边集对应的 JPT
11. create an empty JPT for v_i (the number of rows are 8, the number of columns are 4);
12. **for** each row in the JPT **do**
13. write the existence of edges in the JPT (‘0’ for not exist and ‘1’ for exist);
14. write the value of Pr (the max of probability of exist edges) in the JPT;
15. v_i.*flag*=1;
16. **for** each node v_j in the triangle **do**
17. **if** (|edges of v_j|=2) **then** v_j.*flag*=1;
18. *JPTs*=*JPTs*∪*JPT*;
19. **return** the JPTs.

时间复杂度分析:设不确定图的顶点数为 n,边数为 m.输入过程需要的时间为 $O(n+2m)$,其中,输入边集和概率值分别需要 $O(m)$的时间.设该不确定图中有多条边入射的顶点数为 n_1,且每个顶点的入射边数为 r,则构造这些顶点对应的 JPT 需要的时间为 $O(n_1\times 2^r\times(m_i+1))$.通常情况下,顶点的入射边数是一个较小的常数,因此,构造多入边顶点的 JPT 需要的时间可以化简为 $O(n_1)$.设该不确定图中有 n_2 个三角形,则构造这些三角形对应的 JPT 需要的时间为 $O(n_2)$.因此构造一个不确定图的 JPT 所需要的时间为 $O(n+2m)+O(n_1)+O(n_2)$.由于$(n_1+n_2)\leqslant n$,因此,构造一个不确定图的 JPT 的时间复杂度为 $O(n+2m)$.

4.3 单图上界过滤算法

算法 2 根据单图上界过滤策略判断一个不确定图是否可以进行剪枝.算法的输入为一个不确定图的所有可能世界图 PWG 的概率值以及概率阈值α,输出为是否可以进行剪枝的标记.算法将所有 PWG 的概率值相加,保存于 *sum* 中,并将 *sum* 值与α相比较.若 *sum*<α,则将其进行剪枝.

算法 2. Single graph filtering.

输入:g:an uncertain graph;α:the probability threshold;

输出:whether g should be pruned ('true' for preserve and 'false' for pruning).

1. sum=0;
2. for each PWG of g do //将图 g 所有的 PWG 的概率值相加,存于 sum 中
3. sum+=probability of the PWG;
4. if ($sum<\alpha$) return false; //若 $sum<\alpha$,则将图 g 标记为 false,以便过滤
5. else return true;

时间复杂度分析:设不确定图的可能世界图数为 p,则调用一次单图上界过滤算法对一个不确定图进行过滤判断的时间为 $O(p)$.

4.4 图对上界过滤算法

算法 3 根据图对上界过滤策略判断一个不确定图对是否可以进行剪枝.算法的输入为一个不确定图对的所有可能世界图对的概率值以及概率阈值α,输出为是否可以进行剪枝的标记.算法将所有可能世界图对的概率乘积相加,保存于 sum 中,并将 sum 值与α相比较.若 $sum<\alpha$,则该图对进行剪枝.

算法 3. Graph pair filtering.

输入:(g_i,g_j):a candidate graph pair;α:the probability threshold;

输出:whether (g_i,g_j) should be pruned ('true' for preserve and 'false' for pruning).

1. sum=0;
2. for each g' in PWGs of g_i do //将(g_i,g_j)所有的可能世界图对的概率乘积相加,存于 sum 中
3. for each g'' in PWGs of g_j do
4. sum+=$p(g')\times p(g'')$; //$p(g')$为可能世界图 g'的概率值
5. if ($sum<\alpha$) return false; //若 $sum<\alpha$,则将图对(g_i,g_j)标记为 false,以便过滤
6. else return true;

时间复杂度分析:设不确定图的可能世界图数为 p,则调用一次图对上界过滤算法对一个不确定图对进行过滤判断的时间为 $O(p^2)$.

4.5 顶点过滤算法

算法 4 结合了顶点权值过滤策略和不匹配顶点过滤策略,在计算可能世界图对的相似度过程中对部分可能世界图对进行剪枝.算法的输入是可能世界图对(g',g'')中 g'的所有顶点的权值,其中,g'为图对中顶点数较多的那个可能世界图.算法的输出是(g',g'')是否需要剪枝的标记.算法第 1 行~第 4 行计算 Y 中所有顶点的权值之和,保存于 W_Y 中,若 $W_Y<\gamma$,则可能世界图对(g',g'')可以被剪枝.算法第 5 行~第 8 行计算 YN 中所有顶点的权值之和,保存于 W_{left} 中,若 $W_Y-W_{left}<\gamma$,则(g',g'')也可以被剪枝.

算法 4. Filtering in the *Sim*.

输入:$W(V)$:the weight of vexs in g';γ:the similarity threshold;

输出:whether (g',g'') should be pruned ('true' for preserve and 'false' for pruning).

1. W_Y=0;
2. for each v in Y do //Y 为 g'可在 g''中找到映射的顶点集
3. W_Y+=$W(v)$; //将 Y 中顶点的权值相加,存于 W_Y 中
4. if ($W_Y<\gamma$) return false; //若 $W_Y<\gamma$,则(g',g'')可从计算 Pr 值的所有可能世界图对中过滤掉
5. else W_{left}=0;
6. for each v in YN do //YN 为 Y 中剩余的未找到映射的顶点集
7. W_{left}+=$W(v)$; //将 NY 中顶点的权值相加,存于 W_{left} 中

8. if (W_Y–W_{left}<γ) return false; //若 W_Y–W_{left}<γ,则(g',g'')可从计算 *Pr* 值的所有可能世界图对中过滤掉
9. else return true;

时间复杂度分析:设相似度计算过程中可能世界图 g'对应的 *Y* 的顶点数为|*Y*|,*YN* 的顶点数为|*YN*|,则可能世界图对(g',g'')在相似度计算过程中调用顶点过滤算法进行剪枝的时间为 $O(|Y|+|YN|)$.

4.6 整体相似性连接算法

整个相似性连接方法的伪代码如算法 5 所示.其输入为一个不确定图数据库 *G*,输出为满足相应阈值条件的相似图对的集合.对于 *G* 中的每个不确定图 g_i,算法的第 2 行调用算法 1 读取顶点和边的数据生成 g_i 对应的所有 JPT,第 3 行将这些 JPT 进行连接,生成 g_i 所有的可能世界图,并计算每个可能世界图相应的概率值,第 4 行~第 5 行调用算法 2 计算 g_i 所有可能世界图的概率之和,若和值小于α,则将其过滤,否则,将 g_i 加入第 1 阶段的候选集 *C*1 中.算法的第 6 行~第 8 行调用算法 3 检查 *C*1 中的所有图对(g_i,g_j),若其所有的可能世界图对的概率乘积之和小于α,则将图对(g_i,g_j)过滤,否则,将其加入第 2 阶段的候选集 *C*2 中.算法的第 9 行~第 14 行是对 *C*2 中的所有图对计算 *Pr* 值,从而判断图对是否满足相应阈值下的相似性条件.在计算图对 *Pr* 值的过程中,需要计算该图对中所有可能世界图对的相似度,即 *Sim* 值.根据公式(2)和公式(3),图对的 *Pr* 值为该图对所有 *Sim* 值大于等于γ的可能世界图对的概率乘积之和.若图对的 *Pr* 值大于等于α,则该图对满足相应阈值下的相似性条件,可以作为候选结果之一输出.算法的第 12 行在计算可能世界图对的 *Sim* 值时调用了算法 4 的顶点过滤策略,可以在很大程度上减少真正需要计算的可能世界图对的数量.

算法 5. Similarity join on uncertain graph database.

输入:*G*:an uncertain graph database;

输出:*USJ*:the join result.

1. for each graph g_i in *G* do
2. *JPTs*(g_i)=generate *JPTs*(*V*,*E*,*P*); //调用算法 1 生成每个不确定图的 JPT
3. Join the *JPTs*(g_i) for all the *PWGs*(g_i) and its probability; //将 g_i 的 JPT 进行连接,生成 g_i 的所有可能世界图及其概率
4. *tag*=Single graph filtering(g_i,α); //调用算法 2 对 *G* 中的不确定图进行过滤
5. if (*tag*==“true”) add g_i to *C*1; //若 g_i 不满足过滤条件,则将其加入第 1 阶段候选集 *C*1 中
6. for each graph pair (g_i,g_j), g_i∈*C*1 and g_j∈*C*1 do
7. *tag*=graph pair filtering(g_i,g_j,α); //调用算法 3 对 *C*1 中的所有图对进行过滤
8. if (*tag*==“true”) add (g_i,g_j) to *C*2; //若(g_i,g_j)不满足过滤条件,则将其加入第 2 阶段候选集 *C*2 中
9. for each graph pair (g_i,g_j),(g_i,g_j)∈*C*2 do
10. for each PWG pair (g',g''),g'∈*PWGs*(g_i) and g''∈*PWGs*(g_j) do
11. *Pr*=0;
12. compute *Sim*(g',g'') according to section 4.3; //根据第 4.3 节介绍的方法计算(g',g'')的相似度,计算过程中调用算法 4 直接过滤掉部分 *Sim* 值<γ的可能世界图对,以减少计算量
13. if (*Sim*(g',g'')≥γ) then *Pr*+=*p*(g')×*p*(g''); //根据公式(2)计算不确定图对(g_i,g_j)的 *Pr* 值
14. if (*Pr*≥α) then *USJ*=*USJ*∪(g_i,g_j); //若(g_i,g_j)的 *Pr* 值大于概率阈值α,则将(g_i,g_j)作为一个连接结果添加到 *USJ* 中
15. return *USJ*;

时间复杂度分析:设不确定图数据库 *G* 中的不确定图的数目为 *N*,每个图的顶点数为 *n*,边数为 *m*.

对于 *G* 中的每个不确定图,调用算法 1 生成 JPT 需要的时间为 $O(n+2m)$.设该不确定图生成的 JPT 的数目为 *t*,每个 JPT 的行数为 *r*,将其进行连接计算可能世界图及其概率所需要的时间约为 $O(2^{tr})$.虽然这部分计算的时间复杂度为指数级,但从实验数据来看,大部分的不确定图生成的 JPT 数目都不会太多,因此,实际的计算时间

并没有变得过大.设不确定图的可能世界图的数目为 p,则调用算法 2 对该不确定图进行过滤需要的时间为 $O(p)$.由于 G 中的不确定图的数目为 N,因此这几个步骤的时间复杂度为 $O(N\times(n+2m+2^{tr}+p))$.

设第 1 步过滤得到的候选图数目为$|C1|$,则产生的候选图对数目为$|C1|^2$.调用算法 3 对每个图对进行过滤所需时间为 $O(p^2)$,整个过滤步骤需要的时间为 $O(|C1|^2\times p^2)$.

设第 2 步过滤得到的候选图对数目为$|C2|$.根据第 4.3 节介绍的方法,计算每个候选图对中的每个可能世界图对相似度所需时间为 $O(n\log n+m)$,设调用算法 4 对该候选图对的可能世界图对进行过滤后剩余的可能世界图对为 q,则计算该图对的 Pr 值需要的时间为 $O((n\log n+m)\times q+(|Y|+|YN|)\times p^2)$,其中,$|Y|$表示计算相似度过程中 Y 对应的顶点数,$|YN|$表示 YN 对应的顶点数.最后,根据图对的 Pr 值判断该图对是否为相似性连接的其中一个结果.这个过程总共需要的时间为 $O(|C2|\times q\times(n\log n+m)+|C2|\times p^2\times(|Y|+|YN|))$.

综上所述,整个相似性连接算法的时间复杂度为 $O(N\times(n+2m+2^{tr}+p))+O(|C1|^2\times p^2)+O(|C2|\times q\times(n\log n+m)+|C2|\times p^2\times(|Y|+|YN|))$.

5 实验分析

实验的硬件平台:AMD Athlon™ Ⅱ X4 640 Processor,主频为 3GHZ,4GB 内存.软件平台:Microsoft Windows XP SP3 以及 VC++2005.

在实验中,我们使用了真实的不确定图数据集.该数据集取自 STRING 数据库[10],包含 BioGRID 数据库[11]中的有机物 PPI 网络.PPI 网络中的顶点代表蛋白质,边代表蛋白质间的相互作用,顶点标签为 STRING 数据库提供的蛋白质的 COG 功能注释[9],边的存在概率也是由 STRING 数据库所提供.

我们在 STRING 数据库中随机生成了 5 个不确定图数据库 $G1$,$G2$,$G3$,$G4$ 和 $G5$,各数据库相应的不确定图数分别为 50,100,150,200 和 250.每个不确定图平均包含 385 个节点和 612 条边.每条边平均具有 0.383 的存在概率.

实验的其他参数设置如下:概率阈值α为 0.3~0.5,默认值为 0.4;相似性阈值γ为 0.5~0.7,默认值为 0.6.

由于目前为止在不确定图数据库的相似性连接问题上并没有相关的对比算法,因此,我们的实验主要是对本文所提出的方法及过滤策略进行测试和验证.

首先我们测试本文提出的相似性连接算法的有效性,如图 9 所示,并分析两个参数α和γ对算法运行时间的影响,如图 10、图 11 所示.

图 9 显示了完整的相似性连接算法在不同大小的不确定图数据库中的运行时间,以及使用各阶段过滤策略后算法的运行时间对比,算法运行过程中概率阈值α和相似性阈值γ取默认值.从图中可以看出,随着不确定图数据库的增大,运行时间显著增大,但仍然处于可计算的范畴.由此可见本文提出的相似性连接方法的可行性.从图中还可以看出,本文提出的过滤策略大大缩短了算法的运行时间.其中,使用过滤策略 4.2 即图对上界过滤策略的时间减少得不是很明显,这是因为,该策略需要计算每个图对所有的可能世界图对的概率乘积之和,计算量较大,因此对算法效率的提升贡献相对较小.

图 10 给出了不同的概率阈值下,相似性连接算法在不确定图数据库中的运行时间,算法运行过程中相似性阈值γ取默认值.从实验数据可以看出,随着数据库中不确定图数量的增多,运行时间也逐渐增大.但各数据库运行算法的时间都会随着概率阈值α的增大而减小.这是因为,α值越大,前两个过滤阶段移除的不确定图或者不确定图对越多,下一阶段的计算量就越少,从而使算法的整体运行时间缩短.

图 11 给出了不同的相似性阈值下,相似性连接算法在不确定图数据库中的运行时间,算法运行过程中概率阈值α取默认值.从图中可以看出,在同一个不确定图数据库中,算法的运行时间也随着相似性阈值γ的增大而减小.这是因为,γ值越大,后两个过滤阶段过滤出的可能世界图对就越多,从而使计算相似度算法的运行时间缩短,进而缩短了整体算法的运行时间.

接下来,我们测试本文提出的一组过滤策略的剪枝效率和剪枝能力,如图 12~图 14 所示.

图 12 展示了不确定图数据库 $G1$ 中不同概率阈值和相似性阈值的剪枝时间对比情况.从图中可以看出,剪枝时间随着概率阈值α的增大而缩短,但剪枝时间不会随着相似性阈值γ的变化而有较大幅度的变化,因为γ值不影响

前两步的过滤时间,后两步过滤策略在计算相似度时需要计算所有的可能世界图对,剪枝时间也不会受γ值的影响.

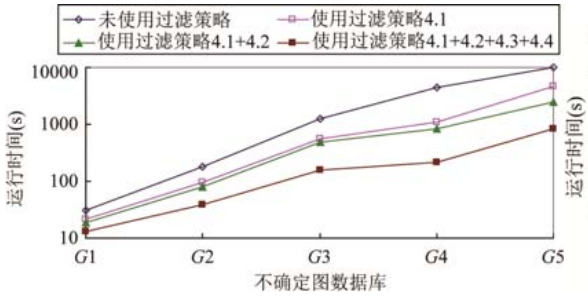

Fig.9 The running time comparison of algorithms

图 9 算法的运行时间对比

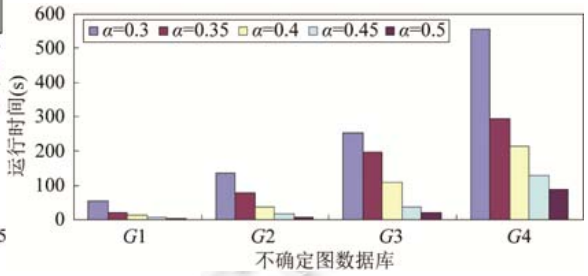

Fig.10 Influence of α on the running time

图 10 α值对算法运行时间的影响

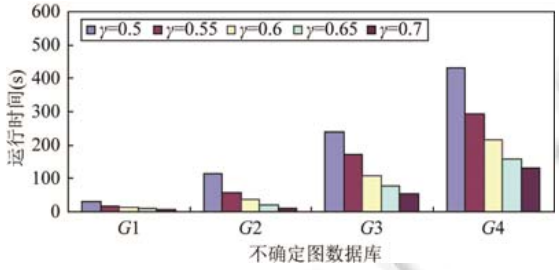

Fig.11 Influence of γ on the running time

图 11 γ值对算法运行时间的影响

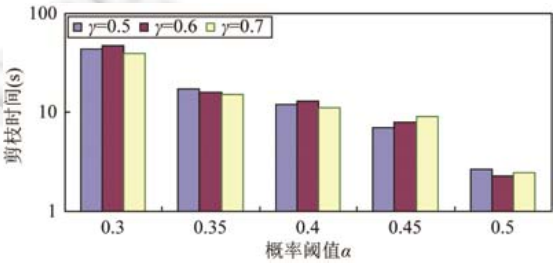

Fig.12 The pruning efficiency of filtering algorithms

图 12 过滤算法的剪枝效率

图 13 分析了前两个过滤策略剪枝前后的候选图对数量对比情况.图中的 *GP* 表示剪枝前数据库中的图对数量,*C*1 表示执行单图上界过滤策略后的候选图所组成的图对数量,*C*2 表示执行图对上界过滤策略后的候选图对数量,算法运行过程中概率阈值和相似性阈值取默认值.从图 13 可以看出,在我们随机生成的 5 个数据库中,单图上界过滤策略可以过滤掉数据库中 1/2 左右的图对,图对上界过滤策略则可以过滤掉 *C*1 中 1/3 左右的图对.虽然这两个过滤策略的实际剪枝能力与数据库中不确定图的概率值有关,但是实验结果表明,这两个过滤策略通常可以达到很好的剪枝效果.由于后两个剪枝策略是针对每个候选图对中的可能世界图对进行的过滤,不能直接与前两个候选图对数相比较,因此没有在图 9 中显示出来.

图 14 给出了后两个过滤策略剪枝前后的可能世界图对数量对比情况,并与文献[7]中相应的过滤策略进行了比较.本次实验采用了上个实验中 5 个数据库剪枝后得到的候选图对来生成相应的可能世界图对,并分别记为 *G*1′,*G*2′,…,*G*5′.图中的 *PWG* 表示剪枝前可能世界图对的数量,*CS* 表示采用文献[7]中的顶点相似性上界过滤策略剪枝后剩余的可能世界图对数量,*VS* 表示采用本文提出的后两种过滤策略即算法 4 中的剪枝方法后剩余的可能世界图对数量.从图 14 可以看出,*CS* 和 *VS* 两种在相似性计算过程中使用的过滤策略都可以对可能世界图对进行有效剪枝,其中,本文提出的过滤策略较文献[7]提出的顶点相似性上界过滤策略具有更好的过滤性能.

最后,我们对算法的查准率和查全率进行分析,如图 15、图 16 所示.所谓的查准率是指在返回的相似图对中正确结果的百分比,查全率是指返回的相似图对在所有正确结果中的百分比.

为了获得不确定图数据库 *G*1~*G*5 中正确的相似图对的数目,我们将数据库中的不确定图通过软件进行转换,使其可以在可视化状态下进行相似性连接的判断,并将这些正确的相似图对和实验结果进行对比,从而计算出本文算法的查准率和查全率.从图中可以看出,本文所提出的方法在查准率和查全率上都具有很高的近似质量,且不会随着概率阈值和相似性阈值的变化而产生较大的波动,其百分比都大于 85%.

从以上测试数据及实验分析可以得出:在不确定图数量较少且图中顶点数和边数较小的情况下,本文所提出的相似性连接算法具有较好的可行性和有效性,算法中所采用的一组过滤策略可以过滤掉很大一部分的不相似图对,具有较好的剪枝能力和剪枝效率,且算法可以达到较高的查询质量.但是随着不确定图数量的增多或者不确定图中顶点数和边数的增大,算法的运行时间会产生较大幅度的增加,因此不适用于大规模的不确定图数据库.

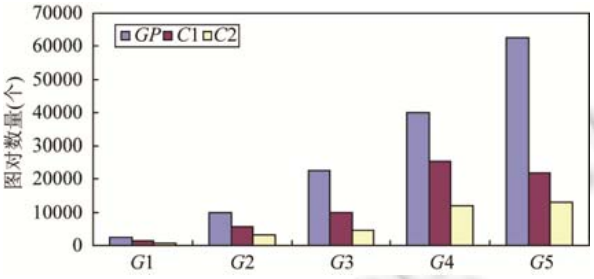

Fig.13 The pruning capability of filtering algorithms (1)

图 13 过滤算法的剪枝能力(1)

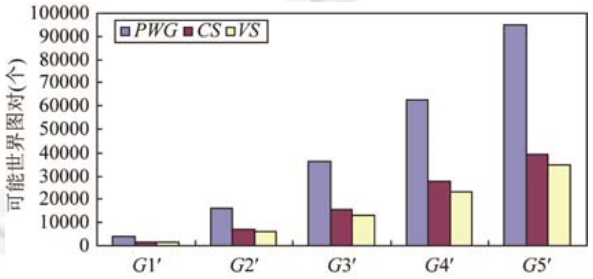

Fig.14 The pruning capability of filtering algorithms (2)

图 14 过滤算法的剪枝能力(2)

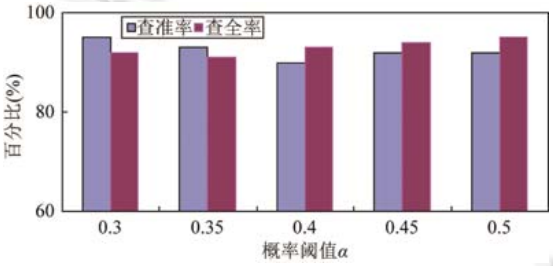

Fig.15 Precision and recall (1)

图 15 查准率和查全率(1)

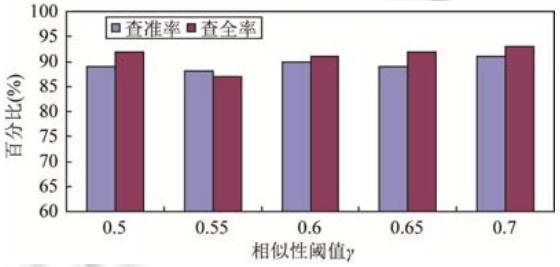

Fig.16 Precision and recall (2)

图 16 查准率和查全率(2)

6 总结和展望

目前大部分对于不确定图的研究都是基于概率独立模型,而实际应用中不确定图的边的存在概率通常是有关联的.本文根据这个特征,选择了一种概率相关模型对不确定图进行建模,将相互影响较大的邻边的概率构造成联合概率分布表 JPT,以此作为不确定图的概率输入.本文在相关文献的启发下,对不确定图数据库上的相似性连接进行了形式化的定义,根据该定义确定了算法的基本步骤并设计了一组过滤策略.实验结果表明,本文提出的方法具有较好的可行性和准确性.

但是,由于本文所讨论问题的复杂性,本文所提出的算法只适用于较小规模的不确定图数据库上的相似性连接.在后续的研究工作中,我们将继续对不确定图数据库上的相似性连接问题进行深入的分析,希望能够提出更高效的处理方式,并借助大数据处理的并行计算框架,进一步研究大规模不确定图数据库上的相似性连接问题.

References:

[1] Yuan Y, Wang GR. Answering probabilistic reachability queries over uncertain graphs. Chinese Journal of Computers, 2010, 33(8):1378–1386 (in Chinese with English abstract).

[2] Jin R, Liu L, Ding B, Wang H. Distance-Constraint reachability computation in uncertain graphs. Proc. of the VLDB Endowment, 2011,4(9):551–562.

[3] Potamias M, Bonchi F, Gionis A, Kollios G. *K*-Nearest neighbors in uncertain graphs. Proc. of the VLDB Endowment, 2010,3(12): 997–1008.

[4] Yuan Y, Wang GR, Chen L, Wang HX. Graph similarity search on large uncertain graph databases. The VLDB Journal, 2015, 24(2):271–296.

[5] Bunke H, Allermann G. Inexact graph matching for structural pattern recognition. Pattern Recognition Letters, 1983,1(4):245–253.

[6] Zhao X, Xiao C, Lin X, Wang W. Efficient graph similarity joins with edit distance constraints. In: Proc. of the IEEE Int'l Conf. on Data Engineering. IEEE, 2012. 834–845.

[7] Wang Y, Wang HZ, Li JZ, Gao H. Efficient subgraph join based on connectivity similarity. Word Wide Web-Internet & Web Information Systems, 2014,18(4):1–17.

[8] Ma YZ, Meng XF. Set similarity join on massive probabilistic data using MapReduce. Distributed & Parallel Databases, 2014, 32(3):447–464.

[9] Biswas S, Morris R. Exor: Opportunistic multi-hop routing for wireless networks. In: ACM SIGCOMM Computer Communication Review. ACM, 2005. 133–144.

[10] Andrew CA, Arnaud C, Montecchi PL, *et al*. Mint: The molecular interaction database. Nucleic Acids Research, 2007,35(D1): D572–4.

[11] Chua HN, Sung WK, Wong L. Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. Bioinformatics, 2006,22(13):452–452.

附中文参考文献:

[1] 袁野,王国仁.面向不确定图的概率可达查询.计算机学报,2010,33(8):1378–1386.

缪丰羽(1983－),女,福建福安人,讲师,主要研究领域为图结构数据处理,大数据分析与管理,XML 数据查询优化.

王宏志(1978－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为数据库,大数据,数据质量.