ISSN 1000-1239/CN 11-1777/TP
49(Suppl.): - , 2012

正文排版44行，双栏，每栏22字

基于 MapReduce 的不确定子图查询处理

题目三号

韩 璐 王朝坤 阮文静 欧晓平 仇 萍

作者四号

(清华大学软件学院 北京 100084)
(清华信息科学与技术国家实验室(筹) 北京 100084)

单位小五号

(信息系统安全教育部重点实验室(清华大学) 北京 100084)
(hanlucc@163.com)

MapReduce Based Uncertain Subgraph Query Processing

英文题目四号

Han Lu, Wang Chaokun, Ruan Wenjing, Ou Xiaoping, and Qiu Ping

姓名五号

(*School of Software*, *Tsinghua University*, *Beijing* 100084)
(*Tsinghua National Laboratory for Information Science and Technology*, *Beijing* 100084)
(*Key Laboratory for Information System Security*, *Ministry of Education*, *Beijing* 100084)

单位小五号

Abstract Graph data structures are highly valued by the expressive capability and widely used in various areas for data modeling. Usually, data cannot be measured accurately in many real-world applications. Uncertain graphs emerge when this kind of data is modeled by graph data structures. In this paper, we address the problem of uncertain subgraph query processing, in which the graph database and the query graph are both uncertain. We propose a MapReduce based algorithm called mutual-match to find out all the same components between two uncertain graphs. An effective index structure is brought forward to accelerate the query processing. Experimental results on the real-world dataset demonstrate the correctness, efficiency and scalability of our approach.

摘要小五号

Key words uncertain graphs; subgraph isomorphism; mutual-match; MapReduce

关键词小五号

摘 要 图数据结构具有较强的模拟复杂结构的能力，能够很好地表达数据对象之间的关联，广泛地用于各领域非结构化数据建模.生产生活中，大量数据由于无法精确地度量而带有不确定性.这样的数据以图数据结构建模就形成了不确定图.在不确定图上的子图查询处理研究领域，存在查询图是不确定图的情况.我们将这种数据图和查询图均为不确定图情况下的查询称为不确定子图查询.为解决这一查询问题，提出了基于 MapReduce 的双向匹配查询算法，并提出了有效的索引结构以提高查询算法的效率.最后，在真实数据集和合成数据集上的实验结果证明了算法的正确性、高效性和扩展性.

摘要小五号

关键词 不确定图；子图同构；双向匹配；MapReduce

关键词小五号

中图法分类号 TP391

中图法小五号

正文五号宋体

随着信息技术的发展，越来越广泛地用于各领域非结构化数据建模，如社会网络、蛋白质交互网、道路交通网等.由于测量手段、数据噪声、抽样统计等因素的限制，数据的不确定性普遍存在.近年来，不确定图数据得到了越来越广泛的重视.

在与不确定图相关的子图查询问题中，查询图也可能是不确定图.例如，在社会网络中，我们得到的成员间影响度模型本身是不确定的.使用这样的模型在社会网络中查询相关信息，即为数据图和查询图均为不确定图情况下的查询问题.我们将此类

收稿日期:2012-07-06
基金项目:国家自然科学基金项目(61170064,60803016);清华信息科学与技术国家实验室(筹)学科交叉基金项目;“核高基”国家科技重大专项基金项目(2010ZX010402-002-002-01)

查询称为“不确定子图查询”.

目前,针对不确定图的研究已涌现出一批成果. 文献[1]研究不确定图上的概率 Top-K 子图匹配问题,提出基于磁盘的索引结构 NG-Index. 文献[2]将子图查询问题扩展到不确定图数据流领域. 文献[3]介绍了在不确定图上执行 K-NN 查询的框架,提出了基于取样的算法. 文献[4]提出用概率统计的方法解决不确定图上的可信赖子图查询问题. 文献[5]提出的 UGRAP 算法在不确定图中挖掘频繁子图模式.

目前,并行程序设计框架 MapReduce[6] 已被应用到图数据查询处理研究领域. 文献[7]利用 MapReduce 查找图中的密集子图. 文献[8]提出了在 MapReduce 上分析 RDF 图的方法. 文献[9]提出了在 MapReduce 上通过并行的广度优先搜索来计算单源最短路径的算法. 但是,不确定图上的 MapReduce 查询处理方法尚未见公开报道.

本文将探讨如何基于 MapReduce 解决不确定子图查询问题. 先前的研究工作都没有涉及这种数据图和查询图均为不确定图的情况. 文献[10]考虑了单机环境中不确定图数据上的不确定查询处理问题,但是当查询图规模较大时,无法支持高效的查询处理. 本文主要贡献为:

1) 设计并实现了基于 MapReduce 的双向匹配查询算法以解决不确定子图查询问题.

2) 将不确定数据图分解为一系列小组件,并基于这些组件构建索引,有效提高了查询处理算法的执行效率.

3) 真实数据和合成数据上的大量实验结果表明了本文所提算法的正确性、高效性和可扩展性.

1 基本概念

一级标题小四黑

定义 1. 不确定图. 一个不确定图 G 是一个五元组 (V,E,Σ,θ,p),其中 V 是顶点集,$E\subseteq V\times V$ 是边集,Σ 是标签集,$\theta:V\to\Sigma$ 是为顶点分配标签的函数,$p:E\to(0,1]$ 是 G 中边 E 的存在概率函数.

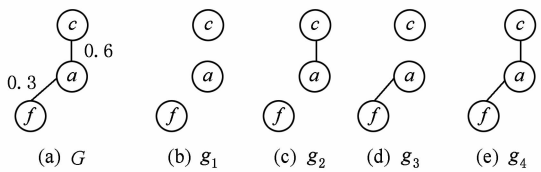
在现有的不确定数据建模方法中,建模核心思想都源自于可能世界模型[11]. 一个不确定图的可能世界是一个由若干确定图组成的集合. 集合中的每个确定图都是不确定图边的组合的一个实例. 不确定图 G 的可能世界表示为 $W_G=\{g_1,g_2,g_3,\cdots,g_n\}$.

如图 1 所示,G 是不确定图. 由于 G 最多有 3 条边,所以 G 的可能世界 W_G 中共包含 2^3 个实例. 每个实例 g_i 有一个存在概率 $Pro(g_i)$,计算公式为

$$Pro(g_i)=\prod_{e\in E_{g_i}}p(e)\prod_{e\in E_G\setminus E_{g_i}}(1-p(e)).$$

以 g_2 为例,可得:

$$Pro(g_2)=0.6\times(1-0.3)=0.42.$$



(a) G (b) g_1 (c) g_2 (d) g_3 (e) g_4

图 1 不确定图 G 和 G 的可能世界 W_G

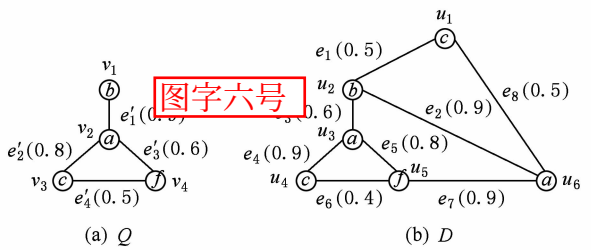
定义 2. 不确定子图查询. 对于不确定查询图 Q 和不确定数据图 D,不确定子图查询返回一系列匹配对. 每个匹配对表示为一个二元组 $m(q,d)$,且满足条件:

1) q 和 d 分别是 Q 和 D 的子图;

2) q 和 d 满足同构条件.

对于不确定图 q 和 d,同构关系表示为双射函数 $f:q\to d$,使得 $f(v')=v$,$f(e')=e$,且 $\theta(v')=\theta(v)$. 其中,$v'\in V_q$,$v\in V_d$,$e'\in E_q$,$e\in E_d$.

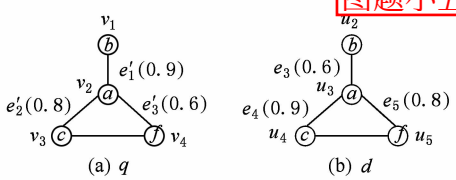
如图 2 所示,若限定同构子图的匹配边数不能小于 3,不确定查询图 Q 和不确定数据图 D 共有 5 个匹配对. 其中一个匹配对 $m(q,d)$ 如图 3 所示.



(a) Q (b) D

图 2 不确定查询图 Q 和不确定数据图 D

图题小五号



(a) q (b) d

图 3 Q 和 D 的一个匹配对 $m(q,d)$

定义 3. 匹配概率. 不确定子图查询匹配对 $m(q,d)$ 的匹配概率表示为:

$$Pro(q,d)=Pro(q)\times Pro(d),$$

其中 $Pro(q)$(或 $Pro(d)$)为 W_Q(或 W_D)中所有包含 q(或 d)的实例的存在概率的和. 以图 3 所示的匹配对 $m(q,d)$ 为例,可得 $Pro(q)=0.9\times0.8\times0.6\times(1-0.5)+0.9\times0.8\times0.6\times0.5=0.9\times0.8\times0.6$,$Pro(q,d)=(0.9\times0.8\times0.6)\times(0.6\times0.9\times0.8)=0.186\,624$.

2　Brute Force 算法

不确定查询图 Q 和不确定数据图 D 的可能世界模型分别为集合 $W_Q=\{g_1,g_2,\cdots,g_s\}$ 和 $W_D=\{g_1,g_2,\cdots,g_t\}$. 为了使不确定的查询图更有现实应用意义,选取 W_Q 中的连通图作为查询图,由集合 $CS(Q)=\{q_1,q_2,\cdots,q_m\}$ 表示. 由于集合 W_D 和 $CS(Q)$ 中的每个元素都可以看作是一个带有概率的确定图,所以不确定子图查询问题可以转化为确定图集 W_D 上的 m 次确定的子图查询处理. 此即为解决不确定子图查询问题的 brute force(BF)算法. 在现有的子图查询索引构造方法中,gIndex[12] 在稀疏图数据集上的查询性能最好[13]. 因此,我们为 W_D 中的确定图建立 gIndex 索引.

W_D 的规模随着 D 边数的增加呈指数倍增长,而且当 Q 较大时,$CS(Q)$ 也会有很大的规模. 除此之外,查询需要多次进行子图同构检测. 所以,BF 算法的时间复杂度和空间复杂度均为指数级,仅适用于查询图和数据图规模都很小的情况下的不确定子图查询.

3　基于 MapReduce 的双向匹配查询算法

由于查询图和数据图都是不确定的,不确定子图查询问题需要在查询过程中根据查询图的结构逐步确定出与之相匹配的数据图的结构,同时,查询图的结构也需要依照数据图的信息确定出来. 这种查询处理是一个查询图与数据图双向匹配的过程. 我们提出了基于 MapReduce 的双向匹配查询(MRMM)算法. 它不需要计算 Q 和 D 的可能世界,也不需要进行同构检测. MRMM 算法将 Q 和 D 分解成一系列“二边图”. 然后,通过边连接操作“EJoin”,在 MapReduce 任务链中将匹配的二边图分别组合到一起,形成 q 和 d. 从而找出所有不确定子图匹配对 $m(q,d)$.

定义 4. 二边图(bi-edge graph). 不确定图 G 的一个二边图定义为 $\lambda(e_l(v_l,v_m),e_r(v_m,v_r))$. λ 是 G 的一个子图,拥有 v_l,v_m 和 v_r 3 个顶点,e_l 和 e_r 两条边. V_m 是该二边图的中间顶点(middle node). 集合 $Bie(G)$ 包含不确定图 G 中所有的二边图.

以图 2 中的 Q 为例,其二边图集合 $Bie(Q)$ 如图 4 所示:

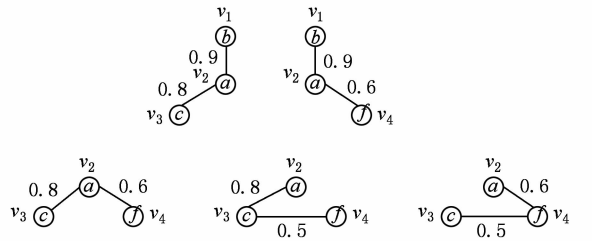


图 4　$Bie(Q)$

定义 5. 边连接操作(EJoin). 给定两个不确定图 G_1,G_2,以及 G_1,G_2 的一条公共边 e,EJoin 操作通过合并公共边 e 将 G_1 和 G_2 连接起来,组成不确定图 G_3.

例如,对于两个不确定图 G_1 和 G_2,以及它们拥有的一条公共边 e_2,$EJoin(G_1,G_2,e_2)$ 将两个不确定图连接后得到不确定图 G_3. 连接过程如图 5 所示:

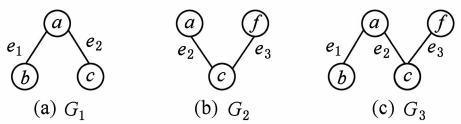


图 5　$EJoin(G_1,G_2,e_2)=G_3$

3.1　基于二边图的索引结构

在查询开始之前,首先为不确定数据图 D 建立索引. 将 D 分解为一系列二边图,得到 $Bie(D)$. 然后,将每一个二边图表示为“标签+值信息”的型式. 对于每一个二边图 λ,$\lambda(e_l(u_l,u_m),e_r(u_m,u_r))\in Bie(D)$,其标签可以表示为“$label_l/label_m/label_r$”,分别为 3 个顶点的标签. $label_m$ 是顶点 u_m 的标签,$label_l$ 和 $label_r$ 满足偏序关系(如按字典序升序排列). λ 的值按照(dbi_id,e_l,e_r,u_m,u_l,u_r)的格式保存. dbi_id 为二边图的唯一标识. e_l,e_r,u_m,u_l,u_r 的排列顺序分别与二边图标签的字符串连接顺序相对应. 最后,将所有的二边图按照标签索引. 索引由哈希链表实现. 表 1 为图 2 的不确定数据图 D 的索引,表 2 为查询图 Q 的索引.

表 1　D 的索引　表题小五号

Key	Value(s)
$a/c/b$	$[1,e_8,e_1,u_1,u_6,u_2]$
$a/b/c$	$[2,e_3,e_1,u_2,u_3,u_1]$,$[3,e_2,e_1,u_2,u_6,u_1]$
$a/b/a$	$[4,e_2,e_3,u_2,u_6,u_3]$
$b/a/f$	$[5,e_2,e_7,u_6,u_2,u_5]$,$[6,e_3,e_5,u_3,u_2,u_5]$
$b/a/c$	$[7,e_2,e_8,u_6,u_2,u_1]$,$[8,e_3,e_4,u_3,u_2,u_4]$
$c/a/f$	$[9,e_4,e_5,u_3,u_4,u_5]$,$[10,e_8,e_7,u_6,u_1,u_5]$
$a/c/f$	$[11,e_4,e_6,u_4,u_3,u_5]$　表字六号
$a/f/a$	$[12,e_5,e_7,u_5,u_3,u_6]$
$a/f/c$	$[13,e_5,e_6,u_5,u_3,u_4]$,$[14,e_7,e_6,u_5,u_6,u_4]$

表 2 Q 的索引

Key	Value(s)
$b/a/c$	$[1,e_1',e_2',v_2,v_1,v_3]$
$b/a/f$	$[2,e_1',e_3',v_2,v_1,v_4]$
$c/a/f$	$[3,e_2',e_3',v_2,v_3,v_4]$
$a/c/f$	$[4,e_2',e_4',v_3,v_2,v_4]$
$a/f/c$	$[5,e_3',e_4',v_4,v_2,v_3]$

3.2 MapReduce 输入文件

对于$\forall\lambda'\in Bie(Q)$,在 D 的索引中找出与λ'标签相同的二边图,将匹配信息写入 MapReduce 输入文件中. 文件格式如图 6 所示:

e_l	e_r	qbi_id	u_m	$posv$	$pose$

图 6 MapReduce 输入文件格式

两个匹配的二边图$\lambda'(e_l'(v_l,v_m),e_r'(v_m,v_r))$和$\lambda(e_l(u_l,u_m),e_r(u_m,u_r))$的匹配信息构成输入文件中的一行记录. 其中$\lambda'\in Bie(Q)$,$\lambda\in Bie(D)$. e_l,e_r,u_m 均来自于λ. qbi_id 是二边图λ'的唯一标识. $posv$ 和 $pose$ 分别是存储匹配顶点和边的向量. 以图 2 中 Q 和 D 为例,可得到如表 3 所示的输入文件.

表 3 Q 和 D 的输入文件 F

e_l	e_r	qbi_id	u_m	$posv$	$pose$
e_2	e_8	1	u_6	$\langle v_1:u_2,v_2:u_6,v_3:u_1\rangle$	$\langle e_1':e_2,e_2':e_8\rangle$
e_3	e_4	1	u_3	$\langle v_1:u_2,v_2:u_3,v_3:u_4\rangle$	$\langle e_1':e_3,e_2':e_4\rangle$
e_2	e_7	2	u_6	$\langle v_1:u_2,v_2:u_6,v_4:u_5\rangle$	$\langle e_1':e_2,e_3':e_7\rangle$
e_3	e_5	2	u_3	$\langle v_1:u_2,v_2:u_3,v_4:u_5\rangle$	$\langle e_1':e_3,e_3':e_5\rangle$
e_4	e_5	3	u_3	$\langle v_3:u_4,v_2:u_3,v_4:u_5\rangle$	$\langle e_2':e_4,e_3':e_5\rangle$
e_8	e_7	3	u_6	$\langle v_3:u_1,v_2:u_6,v_4:u_5\rangle$	$\langle e_2':e_8,e_3':e_7\rangle$
e_4	e_6	4	u_4	$\langle v_2:u_3,v_3:u_4,v_4:u_5\rangle$	$\langle e_2':e_4,e_4':e_6\rangle$
e_5	e_6	5	u_5	$\langle v_2:u_3,v_4:u_5,v_3:u_6\rangle$	$\langle e_3':e_5,e_4':e_6\rangle$
e_6	e_7	5	u_5	$\langle v_2:u_4,v_4:u_5,v_3:u_6\rangle$	$\langle e_3':e_6,e_4':e_7\rangle$

3.3 MapReduce 任务链

在生成输入文件后,MRMM 算法在 MapReduce 框架内通过 $EJoin$ 操作完成匹配对的连接组合. 不确定子图查询匹配的过程由 MapReduce 任务链来实现. 图 7 显示了 MapReduce 任务链的基本工作模型.

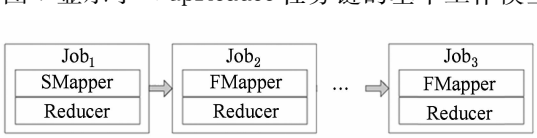


图 7 MapReduce 任务链

在 MapReduce 将输入数据分配给不同的 mapper 后,SMapper 将输入文件中的每一行记录映射为两个 key-value 对. 记录中的前两个元素(e_l 和 e_r)分别作为键值对的 key,剩余的信息作为匹配信息存到键值对中. 例如,表 3 中的第一行记录将生成两个键值对:$\langle e_2;\ e_8,u_6,1,\langle v_1:u_2,v_2:u_6,v_3:u_1\rangle,\langle e_1':e_2,e_2':e_8\rangle\rangle$和$\langle e_8;\ e_2,u_6,1,\langle v_1:u_2,v_2:u_6,v_3:u_1\rangle,\langle e_1':e_2,e_2':e_8\rangle\rangle$.

在图 7 所示的 Job_1 中,$SMapper$ 输出的数据被分配给不同的 $Reducer$. $EJoin$ 的操作即在 $reduce$ 阶段进行. $Reducer$ 在所有键值对中找出键值相同的部分. 在使用 $EJoin$ 进行连接操作之前,$MRMM$ 算法首先要通过若干可连接性检测来判断两个键值对是否是可连接的. 可连接性检测规则包括:1)公共边检测;2)连接方式检测;3)同一性检测.

1) 公共边检测

输入文件中的 qbi_id 数值表明当前用于连接操作的来自不确定数据图的二边图是与不确定查询图中的哪一个二边图相匹配的. 当不确定查询图的两个二边图具有一条公共边时,不确定数据图中分别与之匹配的两个二边图才是可以连接的.

2) 连接方式检测.

通过 EJoin 操作进行连接的二边图有两种连接方式. 如图 8 所示,当两个二边图的中间顶点能够重合,则连接方式为 middle-join,当中间顶点不能重合时,连接方式为 nonmiddle-join.

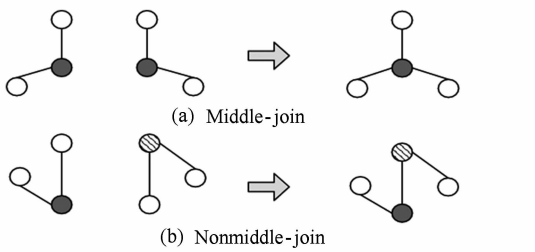


图 8 EJoin 的两种连接方式

对于 Reducer 处理的两个键值对而言,EJoin 连接操作要求当前用于连接的来自不确定查询图的两个二边图的连接方式与不确定数据图中相匹配的两个二边图的连接方式相同.

为了进行前两项检测,定义如下两个函数. 其中$\lambda_1',\lambda_2'\in Bie(Q)$.

$$comf(\lambda_1',\lambda_2')=\begin{cases}0, & \text{if } \lambda_1' \text{ and } \lambda_2' \text{ haven't a common edge;}\\ 1, & \text{if } \lambda_1' \text{ and } \lambda_2' \text{ have a common edge.}\end{cases}$$

$$diref(\lambda_1',\lambda_2')=\begin{cases}0, & \text{if } \lambda_1' \text{ and } \lambda_2' \text{ need a nonmiddle-join;}\\ 1, & \text{if } \lambda_1' \text{ and } \lambda_2' \text{ need a middle-join.}\end{cases}$$

3) 同一性检测.

同一性检测进一步保证了匹配的正确性. 如果两个键值对是可以进行 EJoin 连接的,则不确定查

询图和不确定数据图中的匹配点必须是一一对应的,即来自两个键值对的 $posv$ 中,顶点满足一一映射关系.

当且仅当键值相同的两个键值对通过所有检测时,Reducer 执行 EJoin 操作,将两个键值对组合成一个匹配边数更多的匹配对.假设可连接性参数为 $joinable$,键值相同的两个键值对中当前用于连接的二边图分别是 $\lambda_1',\lambda_2' \in Bie(Q)$,$\lambda_1,\lambda_2 \in Bie(D)$,并且 λ_1,λ_2 分别与 λ_1',λ_2' 相匹配,则可连接性检测依次序如下进行:

1) 计算函数 $comf(\lambda_1',\lambda_2')$的值.如果值为 1,则继续进行剩余检测;如果值为 0,则表明 λ_1 和 λ_2 不可连接,$joinable$=false,结束可连接性检测.

2) 计算函数 $diref(\lambda_1',\lambda_2')$的值.如果值为 1,则继续进行剩余检测;如果值为 0,则表明 λ_1 和 λ_2 不可连接,$joinable$=false,结束可连接性检测.

3) 假设两个键值对中的匹配点向量分别是 $posv_1$ 和 $posv_2$.合并 $posv_1$ 和 $posv_2$ 并验证合并后的向量中,顶点的匹配关系是否满足一一映射关系.如果满足,则 $joinable$=true,结束可连接性检测;如果不满足,则表明 λ_1 和 λ_2 不可连接,$joinable$=false,结束可连接性检测.

如果 $joinable$=true,Reducer 执行 EJoin 操作,将两个键值对连接成一个新的不确定子图匹配对.Reducer 阶段的算法伪代码如下:

算法 1. Reducer.

```
① reduce(String key,Iterator values) {
②   /*key: a common edge*/
③   /*values: a list of initial maching pairs*/
④   for(any two key-value pairs x and y in
      values){
⑤     bool joinable=checkJoinable(x,y);
⑥     if(joinable==true) {
⑦       q'=EJoin(x.q,y.q);
⑧       d'=EJoin(x.d,y.d);
⑨       compute new posv;
⑩       compute new pose;
⑪       compute Pro(q',d');
⑫     /*MinPros is the threshold of Pro*/
⑬     if(Pro(q',d')≥MinPros) {
⑭     create a new matching pair z;
⑮     if(z has matched all query edges)
⑯       Emit('MYM',z);
⑰     else
⑱       Emit(z).   }}}}
```

由于 Reducer 在连接生成新的匹配对后,直接如 SMapper 一样将新匹配对分解成两个键值不同的记录,作为 MapReduce 任务链中其他 Job 的输入文件,所以除 Job_1 之外,剩余 Job 的 FMapper 并未对上一个 Job 的 Reducer 的输入结果进行处理.任务链中每个 Job 的 Reducer 算法均相同.

引理 1. 如果 $EJ_0=EJoin(\lambda_i,\lambda_{i+1},e_i)$,$EJ_1=EJoin(EJ_0,\lambda_{i+2},e_{i+1})$,可以推出

$$EJ_n=EJoin(EJ_{n-1},\lambda_{i+n+1},e_{i+n}).$$

引理 2. 不确定图上的同构定义为一个双射函数 $f:g\to g'$.MapReduce 输入文件的每行记录中,相匹配的两个二边图分别是 λ_i' 和 λ_i,其中 $\lambda_i' \in Bie(Q)$,$\lambda_i \in Bie(D)$,$i=1,2,\cdots$,可以推出 $f(\lambda_i')=\lambda_i$.

定理 1. 对于给定的不确定查询图 Q 和不确定数据图 D,MRMM 算法能返回正确的不确定子图查询结果.

证明. 在 MRMM 算法中,由引理 1 可以得到:

$$q = EJ'_n = EJoin(EJ'_{n-1},\lambda'_{i+n+1},e'_{i+n}),$$

$$d = EJ_n = EJoin(EJ_{n-1},\lambda_{i+n+1},e_{i+n}).$$

假设 e_i 是 λ_i 和 λ_{i+1}的一条公共边,e_i' 是 λ_i' 和 λ_{i+1}'的一条公共边,由引理 2 可以推出 $f(e_i')=e_i$ 且有:

$$f(EJ'_n) = f(EJoin(EJ'_{n-1},\lambda'_{i+n+1},e'_{i+n})) = EJoin(EJ_{n-1},\lambda_{i+n+1},e_{i+n}) = EJ_n$$

由此可以推出 $f(q)=d$.　证毕.

4　实验结果和分析

4.1　实验环境

BF 算法、基于二边图的索引构建算法和 MapReduce 输入文件生成算法运行在单机环境下.该环境运行 Ubuntu10.04 操作系统,硬件参数为 2.4 GHz Intel® Core™ i5 CPU,2 GB 内存.MRMM 算法运行在 Hadoop 分布式框架的 MapReduce 环境中.分布式集群包含 5 台服务器,均运行 Ubuntu 10.04 操作系统,硬件参数为 2.8 GHz Pentium® Dual-Core CPU,3.2 GB 内存.算法均用 Java 实现.

4.2　真实数据集上的实验

实验使用 CAL road network 真实数据集(http://www.cs.fsu.edu/~lifeifei/SpatialDataset.htm).数据集中的不确定图,共包含 21 692 条边,21 047 个顶点.标签集的大小为 L.实验中,为了能够找出不确定查询图和数据图的全部匹配,将

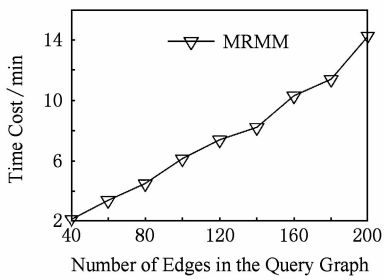
MinPros 的值设为 0.在基于 gIndex 的 BF 算法中,参数 $maxLen=2$.

在实验中,构建 gIndex 索引的时间随着数据量的增长呈指数倍增长趋势.当 $|E_D|=20$ 时将无法为 BF 算法建立 gIndex 索引.而对于拥有少于 20 条边的不确定数据,基于二边图的索引构造时间在 1 ms 之内.

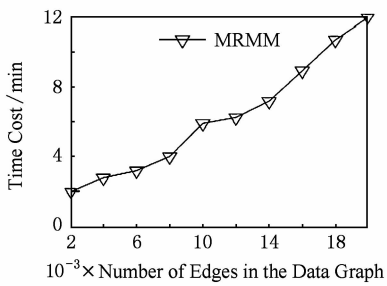
当数据集较小时,MRMM 算法查询效率较低.但随着数据量的增长,MRMM 算法的查询效率会显著优于 BF 算法.

实验验证了 MRMM 算法的扩展性.图 9(a)中,D 共有 21 692 条边,Q 的边数从 40 增至 200.图 9(b)中,Q 有 200 条边,D 的边数从 2 000 增至 20 000.顶点标签集大小 $L=10$.随着不确定查询图和数据图规模的增长,MRMM 算法的查询时间增长趋势平稳.

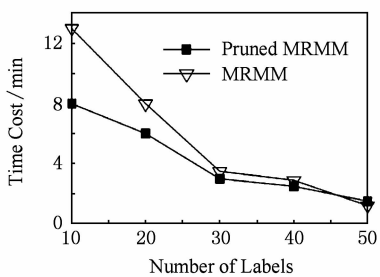
图 9(c)显示了 L 的大小对 MRMM 算法查询效率的影响.实验中,$|E_D|=21,692$,$|E_Q|=200$,L 取值从 10 增加到 50.当 L 值很小时,很多二边图会有相同的标签.因此,需要进行更多的 EJoin 接连操作,增加了算法执行时间.实验还应用了概率剪枝策略.将 *MinPros* 值设为 10^{-10}.结果表明,当顶点标签集的规模很小时,概率剪枝策略会更加有效.这是因为,当 L 值很大时,不确定子图查询匹配的结果会更少.



(a) $|E_Q|$对查询时间的影响



(b) $|E_D|$对查询时间的影响



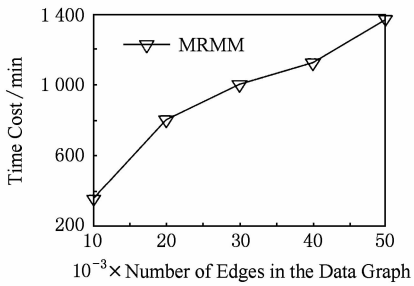
(c) L对查询时间的影响

图 9 真实数据集上的实验结果

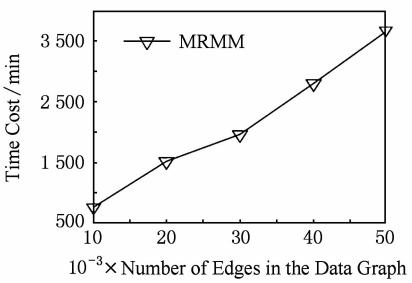
4.3 合成数据集上的实验

实验使用 gengraph_win[14] 合成数据图.数据图共有 50 000 条边,平均度数为 3.标签随机地分配给各个顶点.边上的存在概率服从高斯分布.

由图 10(a)可以看出,在数据量较大的情况下,基于二边图的索引仍然能够在比较理想的时间内构造出来.图 10(b)显示了 D 规模逐渐增大时,生成 MapReduce 输入文件的时间.在合成数据集上也进行了 MRMM 算法的扩展性实验.实验结果与真实数据集上的结论相一致.



(a) 索引构建时间



(b) 输入文件构造时间

图 10 合成数据集上的实验结果

5 结束语

不确定子图查询是不确定图数据库中的一项重要操作.查询图和数据图的不确定性使得不确定子图查询具有很高的时间复杂度.MapReduce 处理大规模数据上的复杂问题的优越性为不确定子图查询问题提供了很好的解决方案.本文提出了 BF 查询算法和基于 MapReduce 的 MRMM 算法,并证明了 MRMM 算法的正确性.真实数据集和合成数据集

上的实验结果表明,MRMM 算法具有很好的查询效率和扩展性.

致谢　感谢彭卓、岳永胜和李瑞对此文的帮助!

参　考　文　献

文献六号

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

[2] Chen L, Wang C. Continuous subgraph pattern search over certain and uncertain graph streams. IEEE TKDE, 2010, 22(8): 1093-1109

[3] Potamias M, Bonchi F, Gionis A, et al. *K*-nearest neighbors in uncertain graphs. PVLDB, 2010, 3(1): 997-1008

[4] Jin R, Liu L, Aggarwal C C. Discovering highly reliable subgraphs in uncertain graphs //Proc of KDD. New York: ACM, 2011: 992-1000

[5] Papapetrou O, Loannou E, Skoutas E. Efficient discovery of frequent subgraph patterns in uncertain graph databases //Proc of EDBT. New York: ACM, 2011: 355-366

[6] Zou L, Chen L, Yu J X, et al. A novel spectral coding in a large graph database //Proc of EDBT. New York: ACM, 2008: 181-192

[7] Dean J, Ghemawat S. MapReduce: Simplified data processing on large clusters. Communications of the ACM. 2008, 51(1): 107-113

[8] Cohen J. Graph twiddling in a MapReduce world. Computing in Science & Engineering. 2009, 11(4): 29-41

[9] Ravindra P, Deshpande V V, Anyanwu K. Towards scalable RDF graph analytics on MapReduce //Proc of Workshop on Massive Data Analytics on the Cloud. New York: ACM, 2010: 1-5

[10] 韩璐,王朝坤,邹鹏,等. 不确定图数据中的不确定查询处理. 计算机研究与发展,2010,47(增刊 1):222-227

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

[12] Yan X, Yu P S, Han J. Graph Indexing: A frequent structure based approach //Proc of SIGMOD. New York: ACM, 2004: 335-346

[13] Han W S, Lee J, Pham M D, et al. iGraph: A framework for comparisons of disk-based graph indexing techniques. PVLDB. 2010, 3(1/2): 449-459

[14] Viger F, Latapy M. Efficient and simple generation of random simple connected graphs with prescribed degree sequence. Computing and Combinatorics, 2005, 3595(2005): 440-449

韩　璐　女,1986 年生,硕士研究生,主要研究方向为图结构数据查询处理.

作者介绍小五号

王朝坤　男,1976 年生,副教授,主要研究方向为音乐数据管理、社交网络与图数据管理.

阮文静　女,1989 年生,硕士研究生,主要研究方向为图结构数据查询处理.

欧晓平　男,1987 年生,硕士研究生,主要研究方向为音乐数据建模以及社会网络.

仇　萍　女,1988 年生,硕士研究生,主要研究方向为基于 ARM 架构的 Java 智能卡的设计与实现.