

基于双向双区间标签实现 k 步可达性查询

宋亚青^{1,2} 武优西^{1,2} 刘靖宇^{1,2} 李 艳³

(河北工业大学计算机科学与软件学院 天津 300401)¹

(河北省大数据重点实验室 天津 300401)² (河北工业大学经济管理学院 天津 300401)³

摘 要 近年来,图的可达性查询已经成为一个研究热点。传统的可达性查询算法——GRAIL 在处理 k 步可达性查询时具有较高的查询效率,但不适合处理不同分支顶点之间的 k 步可达性查询。为了解决上述问题,提出了一种新的双向双区间标签索引,进而实现了 RE-GRAIL 算法,从而有效解决了 k 步可达性查询问题。最后,在 5 个不同特征的数据集上进行实验,并从索引构建时间、索引大小、查询时间、扩展性 4 个方面进行验证。实验结果表明,与众多同类算法相比,RE-GRAIL 算法具有更好的性能。

关键词 k 步可达性查询,不同分支,双向,双区间

中图分类号 TP311 文献标识码 A DOI 10.11896/j.issn.1002-137X.2018.03.028

k -step Reachability Queries Based on Bidirectional Double Interval Labeling Indexes

SONG Ya-qing^{1,2} WU You-xi^{1,2} LIU Jing-yu^{1,2} LI Yan³

(School of Computer Science and Engineering, Hebei University of Technology, Tianjin 300401, China)¹

(Hebei Province Key Laboratory of Big Data Calculation, Tianjin 300401, China)²

(School of Economics and Management, Hebei University of Technology, Tianjin 300401, China)³

Abstract Recently, reachability query is one of the main research topics on graph data. GRAIL can deal with k -step reachability queries efficiently, however, it is not suitable for processing the query in which the vertex pairs are located in different branches. This paper further proposed RE-GRAIL algorithm which employs a bidirectional double interval labeling indexes to tackle the problem. At last, five real datasets were employed to validate the performances of the proposed algorithm in terms of different metrics, including indexing time index size, query processing time and scalability. Experimental results show that RE-GRAIL has better performance than other competitive algorithms.

Keywords Reachability queries within k steps, Different branches, Bidirectional, Double interval

1 序言

随着生物信息学、数据挖掘、异构信息集成系统、社交网络、云计算、交通网络等新兴技术的广泛应用^[1-3], 现实社会中涌现出的数据变得越来越复杂^[4-5]。由于树数据结构模型继承的单一性,图^[6]作为一种能描述复杂结构化的通用数据结构模型,将代替树数据结构模型来实现对复杂数据的存储和分析,而针对大图数据的高效检索问题已成为了当前科学研究所面临的主要难题,同时也成为了研究者广泛关注的热点问题^[7]。

可达性查询是图数据管理中频繁使用的基础操作,并且在现阶段已有较为深入的研究^[8],但对于一些现实问题,仅仅知道两顶点是否可达并不能满足用户的需求。与可达性查询

相比, k 步可达性查询能为用户提供更多的有用信息^[9]。例如:在文献数据库检索系统中若要预判断两篇文献之间引用的紧密程度,可将该系统抽象成一个有向无环图,将一篇论文视为图中的一个顶点,而将论文与参考文献之间的引用关系表示成有向边。基于该模型构造出了引文网络。可达性查询仅能判断出两篇文献是否存在联系,而文献之间的关系紧密程度的求解是基于 k 步可达性查询来实现的。研究者发现,高效地处理有向无环图中的 k 步可达性查询问题在理论研究和实际应用中都起着至关重要的作用^[10-12],从而引发了学者们的关注。周军锋等人^[12]在 GRAIL 算法^[10]的基础上提出了 BiRch 算法,该算法借鉴 GRAIL 算法中提出的单向双区间标签索引,首先过滤掉部分不可达点对,然后对于不能判别的节点,采用递归的方式从顶点对中出度或入度较小的方向

到稿日期:2016-12-06 返修日期:2017-04-15 本文受国家自然科学基金(61673159),河北省自然科学基金(F2016202145),黑龙江省自然科学基金(F2017019),河北省科技计划项目(15210325),河北省教育厅青年基金(QN2014192)资助。

宋亚青(1991—),女,硕士生,主要研究方向为智能计算,E-mail:1198853275@qq.com;武优西(1974—),男,博士,教授,CCF 高级会员,主要研究方向为数据挖掘、智能计算,E-mail:wuc567@163.com(通信作者);刘靖宇(1976—),男,博士,副教授,主要研究方向为数据挖掘、绿色存储;李 艳(1975—),女,博士,副教授,主要研究方向为数据挖掘。

向另外一方查找。显然这是一种贪心策略,难以保障总体访问顶点的数量最小,并且会造成大量节点对的冗余判断。此外,有些研究成果亦可用作 k 步可达性查询,李艳等人^[13-14]分别于 2012 年和 2015 年采用新型数据结构网树^[15-16]研究了两点间存在多少条不同的 k 步路径和两点间存在多少条互不相交的 k 步路径。显然在上述两种研究中,当结果的路径条数大于 1 时,均可表明两点间是 k 步可达的。

当前 k 步可达性查询多是通过构建索引来实现的^[10-12]。但索引的构建和查询效率之间有着密不可分的关系,如果构建的索引越复杂,空间消耗越大,那么查询时间就会越短;相反,构建的索引越简单,空间消耗越小,查询时间就会越长。如何合理地构建索引并得到较高的查询效率成为了研究的热点和难点^[10-12,17-18]。本文正是从此角度出发进行研究。

本文第 2 节给出了问题描述并提出了 RE-GRAIL 算法,通过实例阐明了该算法的工作原理;第 3 节展示了实验结果,验证了本文方法的有效性;最后总结全文。

2 问题描述及求解算法

本文问题可以描述为:给定有向无环图 G , u 和 w 为图 G 中的任意两个顶点,若两顶点之间存在一条长度小于或等于 k 的路径,则顶点 u 在 k 步之内可达顶点 w ,用 $u \rightarrow_k w$ 来表示;否则两顶点在 k 步之内不可达。此外, $inN(u, G) = \{v \mid (v, u) \in E\}$ 表示指向顶点 u 的顶点集,即 u 的入度顶点集; $outN(u, G) = \{v \mid (u, v) \in E\}$ 表示顶点 u 指向的顶点集,即 u 的出度顶点集; $indeg(u, G) = |inN(u, G)|$ 表示顶点 u 的入度; $outdeg(u, G) = |outN(u, G)|$ 表示顶点 u 的出度。

2.1 GRAIL 算法

GRAIL 算法是基于单向双区间标签索引实现的,建立双区间标签的基本思想是:给定有向无环图 G 按照后根遍历顺序时顶点的访问次序为每一个顶点 v 产生两个区间标签,记作 $L_v^1 = [S_v^1, T_v^1]$ 和 $L_v^2 = [S_v^2, T_v^2]$,其中 T_v^1 和 T_v^2 代表区间的终止点值,其值分别是按照从左到右和从右到左的顺序后根遍历图 G 中顶点 v 的访问顺序; S_v^1 和 S_v^2 代表区间的起始点值,其值分别依据 T_v^1 和 T_v^2 生成,具体计算方法如下:

$$S_v^i = \begin{cases} T_v^i, & outdeg(v) = 0 \\ \min\{S_x^i \mid x \in outN(v, G)\}, & \text{其他} \end{cases} \quad (1)$$

其中, $i = 1, 2$ 。

为了便于说明,图 1 给出了有向无环图 G 及其对应的单向双区间标签。图 1 中顶点的从左到右的后根遍历次序为: 8, 2, 15, 7, 6, ...。因此顶点 8 为该次序下第 1 个访问顶点,故 $T_8^1 = 1$ 。当顶点的出度为 0 时,该点的 S_v^2 的值等于 T_v^2 。已知顶点 8 的出度为 0,因此 $S_8^2 = T_8^2 = 1$,故顶点 8 的第一个区间标签为 $[1, 1]$ 。若顶点的出度不为 0,则 S_v^2 为该顶点的所有孩子顶点的 S_x^2 所取的最小值,例如顶点 9 指向顶点 10, 15 和 12,故 $S_9^2 = \min\{S_{10}^2, S_{15}^2, S_{12}^2\} = \min\{1, 3, 3\} = 1$;又因顶点 9 是该访问次序中第 15 个访问的顶点,故顶点 9 的第一个区间标签为 $[1, 15]$ 。同理可生成各个顶点的第二区间标签,这里不再赘述。

GRAIL 算法的 k 步可达性查询处理的过程是:判断顶点

u 到顶点 v 是否 k 步可达,当两个区间标签满足关系 $(L_u^1 \subset L_v^1) \vee (L_u^2 \subset L_v^2)$ 时,直接得出两顶点 k 步不可达的结论;否则需要递归判断 u 的所有孩子顶点与顶点 v 的可达性关系。如果最后判断出顶点的所有孩子顶点都不可达顶点 v ,那么得出结论:顶点 u 到顶点 v 是 k 步不可达的。

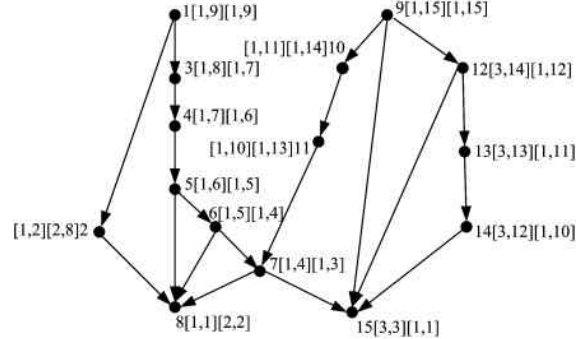


图 1 有向无环图 G 及单向双区间标签

Fig. 1 Directed acyclic graph G and its directional interval labeling indexes

2.2 RE-GRAIL 算法

当数据集中有较多分支,即入度为 0 的顶点比较多时,GRAIL 算法不适合处理不同分支顶点之间的 k 步可达性查询问题。下面举例说明:如图 1 所示的查询 $9 \rightarrow_4 2$,根据单向双区间标签不能快速判断出这两个顶点是 4 步不可达的,由于这两个顶点位于不同的分支,并且两个顶点可以到达一个共同的顶点,这时只能通过递归查询来判断。而递归又会带来大量的时间开销和空间开销,从而严重影响算法的查询效率。为了解决上述问题,本文提出了一种新的区间标签索引方法——双向双区间标签索引。本文提出的 RE-GRAIL 算法则是在新索引上实现的 k 步可达性查询。

2.2.1 基于双向双区间构建索引

给定有向无环图 G ,为图中的每个顶点附加两个区间标签,记作 $L_v^1 = [S_v^1, T_v^1]$ 和 $L_v^2 = [S_v^2, T_v^2]$,其中第一个区间标签 L_v^1 的求解方法与单向双区间标签中 L_v^1 的求解方法相同。 L_v^2 的求解方法是:将按照从左到右的顺序逆向遍历顶点 v 的访问顺序作为 T_v^2 的值,而 S_v^2 的计算公式为:

$$S_v^2 = \begin{cases} T_v^2, & indeg(v) = 0 \\ \min\{S_x^2 \mid x \in inN(v, G)\}, & \text{其他} \end{cases} \quad (2)$$

显然, S_v^2 与 S_v^1 的取值方法截然不同,当顶点 v 的入度为 0 时, S_v^2 的取值等于 T_v^2 ;当顶点 v 的入度不为 0 时,取其对应的最小值。图 1 的有向无环图 G 及其双向双区间标签结果如图 2 所示。由于从左到右的逆向遍历次序为: 1, 2, 3, 4, 5, 6, 9, 7, 8, ...。因此顶点 9 在上述序列中的次序为第 7 个访问顶点。又由于顶点 9 的入度为 0,则 $S_9^2 = T_9^2 = 7$,因此顶点 9 对应的第二个区间标签为 $[7, 7]$ 。若顶点的入度不为 0,则其 S_v^2 为该顶点的所有双亲顶点的 S_x^2 取最小值。以顶点 7 为例进行说明:指向顶点 7 的顶点分别是 6 和 11,因此 $S_7^2 = \min\{S_6^2, S_{11}^2\} = \min\{1, 7\} = 1$;又由于顶点 7 在上述逆向遍历时的次序为第 10 个顶点,因此 7 对应的第二个区间标签为 $[1, 10]$ 。其他顶点的第二个区间标签也可以按照上述原理依

次获得,这里不再赘述。

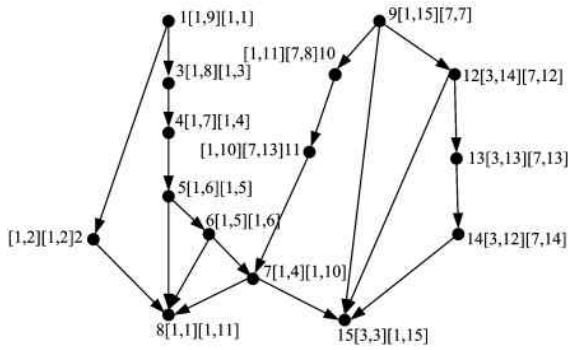


图 2 有向无环图 G 及双向双区间标签

Fig. 2 Directed acyclic graph G and its bidirectional interval labeling indexes

RE-GRAIL 算法判断的依据是:首先考虑第一个区间标签, T_u^1 为从根顶点开始按照从左到右的顺序后根遍历图 G 时顶点 u 的访问顺序,由于顶点 u 在 k 步之内可达顶点 v ,则顶点 u 为顶点 v 的祖先顶点,顶点 v 先于顶点 u 被访问,因此有 $S_v^1 \leq S_u^1$ 且 $T_u^1 > T_v^1$. 故可得出结论 $(L_v^1 = [S_v^1, T_v^1]) \subseteq (L_u^1 = [S_u^1, T_u^1])$. 其次,考虑第二个区间标签, T_u^2 为图 G 从左到右的顺序逆向遍历时顶点 u 的访问顺序,同样由于顶点 u 在 k 步之内可达顶点 v ,则顶点 u 为顶点 v 的祖先顶点,因此逆向遍历图 G 时,顶点 u 一定先于顶点 v 被访问,则有 $S_v^2 \leq S_u^2$ 且 $T_v^2 > T_u^2$. 故可得出结论 $(L_v^2 = [S_v^2, T_v^2]) \subseteq (L_u^2 = [S_u^2, T_u^2])$. 如果这两个区间包含关系中有一个不成立,则可直接得出两顶点对不可达的结论。

2.2.2 算法描述

依据上述原理,RE-GRAIL 算法的伪代码如下。

算法 1 RE-GRAIL 算法//k 步可达性查询 Reachability

(u, v, G, k)

输入:查询顶点对,有向无环图,k(步数)值

输出:构造索引的大小,构造索引的时间,查询时间

```

1. if  $(L_v^1 \not\subseteq L_u^1) \vee (L_v^2 \not\subseteq L_u^2)$  THEN
2.   return False; // 顶点 u 不可达顶点 v
3. ELSE
4.   FOREACH  $c \in \text{Children}(u)$  such that  $(L_v^1 \subseteq L_u^1) \cap (L_v^2 \subseteq L_u^2)$ 
   do
5.     IF Reachability(c, v, G, k-1) THEN
6.       return True; // 顶点  $u \rightarrow_k v$ 
7. return false.
```

用 GRAIL 算法和 RE-GRAIL 算法分别实现查询 $9 \rightarrow_5 5$ 。

1)GRAIL 算法:首先根据图 1 中的单向双区间标签可得顶点 9 的区间标签为 $[1, 15][1, 15]$, 顶点 5 的区间标签为 $[1, 6][1, 5]$, 由于 $[1, 6] \subseteq [1, 15]$ 且 $[1, 5] \subseteq [1, 15]$, 因此不能根据单向双区间标签快速判断出两顶点的可达性。后续判断需要通过单向递归完成,这会严重影响算法的查询效率。

2)RE-GRAIL:根据图 2 中的双向双区间标签索引可得,顶点 9 对应的区间标签为 $[1, 15][7, 7]$, 而顶点 5 对应的区间标签为 $[1, 6][1, 5]$ 。由于两顶点的第二个区间标签满足关系 $[7, 7] \not\subseteq [1, 5]$, 根据前面的原理可直接得出结论 $9 \rightarrow_5 5$ 不成

立,因此不需要通过单向递归来进一步判断,从而大大地提高了算法的查询效率。

3 实验结果与分析

3.1 实验数据集开发环境

本文实验所用的计算机硬件环境是: Intel(R) Core(TM) i7-4790 处理器, 3.60GHz 主频, 8.0GB 内存; 软件环境是: 64 位 Windows7 操作系统, Microsoft Visual Studio 2013 编程环境。本实验使用 C++ 编程语言实现了 PLL^[11], GRAIL^[10], BiRch^[12] 和 RE-GRAIL 共 4 种算法, 其中 PLL 算法根据最短距离来解决 k 步可达性查询问题。

为了准确地比较 4 种算法的查询效率, 实验中对每个数据集随机产生 10 万个顶点对。查询时间(查询处理 10 万个顶点对的总时间)是算法执行 10 次的平均值。

本文实验用到了 5 个数据集, 分别为 Pubmed (pubmed-central. nih. gov), Citeseer (citeseer. ist. psu. edu), Cit-patent^[15-16], Uniprot22m (www. uniprot. org), Ba10k5d。其中 Pubmed 是一个小规模的数据集; Citeseer, Cit-patent 和 Uniprot22m 是 3 个大规模的数据集; Ba10k5d 是随机生成的数据集, 包含 1 万个顶点且顶点的平均度为 5。所有数据集中均含有较多入度为 0 的顶点, 其相关统计信息如表 1 所列, 其中 $|V|$ 表示有向无环图中顶点的个数, $|E|$ 表示边的条数, $indeg_{\max}$ 表示有向无环图中所有顶点的最大入度, $outdeg_{\max}$ 表示所有顶点的最大出度, u 表示根顶点的个数。

表 1 数据集信息

Table 1 Information of datasets

数据集	$ V $	$ E $	$indeg_{\max}$	$outdeg_{\max}$	u
Pubmed	9000	40028	432	87	2609
Citeseer	693947	312282	55757	1	613497
Cit-patent	3774768	16518947	779	770	515785
Uniprot22m	1595444	1595442	1539898	1	1556158
Ba10k5d	10000	49975	501	5	2865

3.2 构造索引大小及时间

表 2 列出了 4 种算法所需构造的索引大小。

表 2 索引大小

Table 2 Size of indexes

算法	Pubmed	Citeseer	Cit-patent	Uniprot22m	Ba10k5d
PLL	1.25	70.36	—	—	3.24
GRAIL	0.19	20.32	116.68	35.44	0.17
BiRch	0.19	20.32	116.68	35.44	0.17
RE-GRAIL	0.18	20.72	114.72	37.66	0.19

(单位: MB)

观察表 2 发现: 由于 GRAIL 和 RE-GRAIL 都是双区间标签索引, 因此其标签索引的大小相差不大, 都比 PLL 构造的索引小。PLL 算法由于适用于小规模数据集, 因此在 Cit-patent 和 Uniprot22m 数据集上未能构建索引, 在表 2 中用“—”表示。BiRch 算法是基于 GRAIL 中的单向双区间标签索引实现的, 因此两种算法构建的索引大小相等。

表 3 列出了 4 种算法构造索引的时间。观察表 3 发现: GRAIL 和 RE-GRAIL 算法构造的时间相差不大, 在其他数据集上, PLL 算法对索引的构造时间均多于 GRAIL 和 RE-GRAIL 算法对索引的构造时间。

表 3 索引构造时间

Table 3 Time of constructing index

(单位:ms)

算法	Pubmed	Citeseer	Cit-patent	Uniprot22m	Ba10k5d
PLL	517.63	2509.18	—	—	19.43
GRAIL	11.70	876.58	873.99	1436.42	13.16
BiRch	12.34	868.58	867.89	1478.16	14.27
RE-GRAIL	14.67	892.18	854.32	1734.71	14.19

3.3 查询时间

表 4 列出了当 $k=5$ 时 4 种算法在不同数据集上的查询时间。观察表 4 发现:与 PLL 算法相比,GRAIL 算法在解决 k 步可达性查询时有较高的查询效率;当数据规模较小时,BiRch 有较高的查询效率;但是对于大规模的数据集,本文提出的 RE-GRAIL 算法具有较高的查询效率。

表 4 查询时间

Table 4 Query time

(单位:ms)

算法	Pubmed	Citeseer	Cit-patent	Uniprot22m	Ba10k5d
PLL	517.63	509.18	—	—	2102.62
GRAIL	63.57	22.87	39.19	59.43	230.51
BiRch	49.76	22.03	53.17	56.42	227.32
RE-GRAIL	59.24	19.62	34.19	54.27	217.93

3.4 不同 k 值的查询处理性能

表 5 列出了采用 RE-GRAIL 算法时在不同特征的数据集上不同 k 值对应的查询时间。观察表 5 可发现: k 值变化对查询时间的影响不大。

表 5 不同 k 值时的查询时间Table 5 Comparison of query time under different k

(单位:ms)

数据集	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$	$k=8$	$k=9$
Pubmed	60.13	61.34	59.10	58.34	60.89	61.23	59.78
Citeseer	19.16	18.27	18.87	16.47	18.56	19.91	17.30
Cit-patent	34.34	37.12	36.89	36.32	35.97	36.63	35.46
Uniprot22m	57.41	54.73	56.24	54.74	54.82	55.53	57.69
Ba10k5d	210.65	214.19	219.77	212.81	220.95	218.28	219.19

结束语 本文在 GRAIL 算法的基础上,提出了基于双向双区间标签索引的优化算法 RE-GRAIL。为了验证该算法的性能,在 5 个不同特征的数据集上进行实验,并从索引构建时间、索引大小、查询时间和可扩展性 4 个方面进行实验。实验结果表明,与其他同类算法相比,RE-GRAIL 算法具有更好的性能。

参 考 文 献

[1] JIANG H,ZHANG J,MA H,et al. Mining authorship characteristics in bug repositories[J]. Science China Information Sciences, 2017,60(1):1-16.

[2] YIN D,GAO H,ZOU ZN,et al. Reachability query in heterogeneous information networks[J]. Journal of Computer Research and Development,2016,53(2):479-491.

[3] HAN Y J. Method for finding critical paths based on concurrent reachable marking graph with Tags [J]. Computer Science, 2016,43(11):121-125. (in Chinese)
韩耀军. 基于带标记的并发可达标识图的关键路径的求解方法[J]. 计算机科学,2016,43(11):121-125.

[4] NIE L,JIANG H,REN Z,et al. Query expansion based on crowd knowledge for code search[J]. IEEE Transactions on

Services Computing,2016,9(5):771-783.

- [5] XUAN J,JIANG H,HU Y,et al. Towards effective bug triage with software data reduction techniques[J]. IEEE Transactions on Knowledge and Data Engineering,2015,27(1):264-280.
- [6] HOU R,WU J G. Efficient reduction algorithms for directed acyclic graph [J]. Computer Science,2015,42(7):78-84. (in Chinese)
侯睿,武继刚. 有向无环图的高效归约算法[J]. 计算机科学,2015,42(7):78-84.
- [7] JIANG J X,YI P P,CHOI B,et al. Privacy-preserving reachability query services for massive networks[C]//25th ACM International Conference on Information and Knowledge Management. Indiana,USA,2016:145-154.
- [8] COHEN E,HALPERIN,KAPLAN H,et al. Reachability and distance queries via 2-hop labels [C]//Proceedings of the 13th ACM-SIAM Symposium on Discrete Algorithms Society for Industrial and Applied Mathematics. New York,USA,2002:937-946.
- [9] FU L Z,MENG X F. Reachability indexing for large-scale graphs:studies and forecasts[J]. Journal of Computer Research and Development,2015,52(1):116-129. (in Chinese)
富丽贞,孟小峰. 大规模图数据可达性索引技术:现状与展望[J]. 计算机研究与发展,2015,52(1):116-129.
- [10] YILDIRIM H,CHAOJI V,ZAKI M J. Grail:Scalable reachability index for large graphs[J]. PVLDB Journal,2010,3(1):276-284.
- [11] AKIBA T,IWATA Y,YOSHIDA Y. Fast exact shortest-path distance queries on large networks by pruned landmark labeling [C]// Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data. New York,USA,2013:349-360.
- [12] ZHOU J F,CHEN W,FEI C P,et al. BiRch:A bidirectional search algorithm for k-step reachability queries [J]. Journal on Communications,2015,16(8):50-60. (in Chinese)
周军锋,陈伟,费春苹,等. BiRch:一种处理 k 步可达性查询的双向搜索算法[J]. 通信学报,2015,16(8):50-60.
- [13] LI Y,SUN L,ZHU H Z,et al. A nettree for simple paths with length constraint and longest path in directed acyclic graph [J]. Chinese Journal of Computer,2012,35(10):2194-2203. (in Chinese)
李艳,孙乐,朱怀忠,等. 网树求解有向无环图中具有长度约束的简单路径和最长路径问题[J]. 计算机学报,2012,35(10):2194-2203.
- [14] LI Y,WU Y X,HUANG C P,et al. Nettoree for maximum disjoint paths with length constraint in DAG [J]. Journal on Communications,2015,36(8):38-49. (in Chinese)
李艳,武优西,黄春萍,等. 网树求解有向无环图中具有长度约束的最大不相交路径 [J]. 通信学报,2015,36(8):38-49.
- [15] WU Y,WANG L,REN J,et al. Mining sequential patterns with periodic wildcard gaps [J]. Applied Intelligence,2014,41(1):99-116.
- [16] WU Y,SHEN C,JIANG H,et al. Strict pattern matching under non-overlapping condition [J]. Science China Information Sciences,2017,60(1):1-16.
- [17] CHENG J,SHANG Z,CHENG H. K-reach:Who is in your small world[J]. PVLDB Journal,2012,5(11):1292-1303.
- [18] CHENG J,SHANG Z,CHENG H. Efficient processing of k-hop reachability queries[J]. VLDB Journal,2014,23(2):227-252.