时序优先级约束的时序模式图强模拟匹配

金浩宇,霍　宏,方　涛
(上海交通大学 电子信息与电气工程学院,上海 200240)

摘　要:图模式匹配是一种在图数据上进行高效查询的重要方法,有着广泛的应用前景,例如知识发现、社交网络分析、智能问答等。大多数现有的研究工作都是基于静态的图数据,而现实生活中的图数据很多属于包含时间信息的时态图,针对时态图上的模式图匹配,该文提出了一种时序优先级约束的时序模式图强模拟匹配算法(Temporal Priority Constrained Graph Pattern Strong Simulation Matching,TPC-GPSSM)。该算法在模式图的图拓扑结构的匹配过程中加入时间顺序约束,即考虑了时态图中不同时态边之间的时序优先级,同时通过设置冗余顶点过滤规则来缩小搜索范围,优化时序检查的队列顺序,以达到提前剪枝、减少计算复杂度的目的。提出了时态边聚合度来评价算法对时态边的过滤效果,在三个时序数据集上的大量实验表明,相比传统的强模拟算法,所提算法能够有效过滤错误结果,并且在不同规模的数据图上均具有良好的性能表现。

关键词:模式图匹配;时态图;强模拟;图模拟;时序模式图
中图分类号:TP301　　**文献标识码**:A　　**文章编号**:1673-629X(2023)06-0088-07
doi:10.3969/j.issn.1673-629X.2023.06.014

Temporal Priority Constrained Graph Pattern Strong Simulation Matching

JIN Hao-yu,HUO Hong,FANG Tao
(School of Electronic Information and Electrical Engineering,Shanghai Jiao Tong University,
Shanghai 200240,China)

Abstract:Graph pattern matching is an important method for efficient query on graph data,which has wide application prospects,such as knowledge discovery,intelligent question answering,social network analysis,and so on. Most of the existing researches are generally based on static graph data. However,many graph data with time information in real world belong to temporal graphs. Aiming at graph pattern matching in temporal graphs,a temporal priority constrained graph pattern strong simulation matching method is proposed. It introduces the constraints of time orders into the pattern graph matching as while matching the graph topology of pattern graphs,namely,it considers the temporal priorities of different temporal edges in temporal graphs. Meanwhile,redundant vertex filtering rules are set to narrow the search scope and optimize the queue of time order,so as to prune the graph in advance and reduce the computational complexity. Moreover,the temporal edge closeness is proposed to evaluate the algorithm's performance by the filtering effects on temporal edges. Experiments results on three temporal datasets have shown that the proposed method can effectively filter out the error matching results compared with the traditional strong simulation algorithm,and also has satisfactory performance on data graphs of different scales.

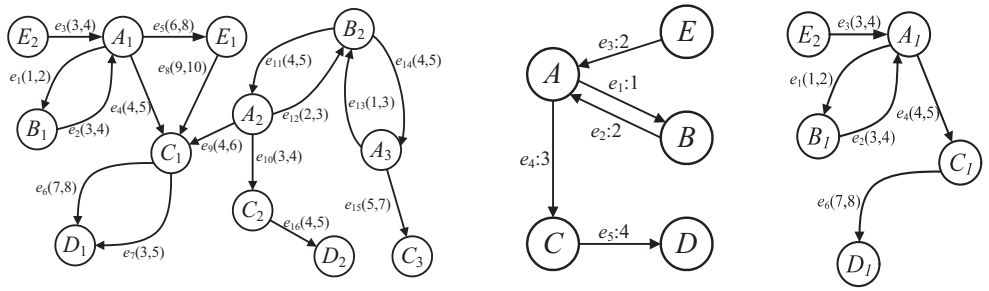
Key words:graph pattern matching;temporal graph;strong simulation;graph simulation;temporal graph pattern

0　引　言

图模式匹配是指在图数据中搜索与待查询模式图相同或相似的子图的一类算法,是实现图数据高效率查询的重要手段,已经应用于众多的领域。例如社交网络分析[1]、知识分析[2]等。现有的图模式匹配方法大多仅适用于静态图,不适用于在现实世界中广泛存在的时态图[3-4],即图数据中的两个顶点之间的边具有时间属性,其记录了两个顶点之间关系的开始与结束时间。图 1(a)给出了一个任务合作时序图,图中每个顶点表示一个人,顶点标签 A,B,C,D,E 表示不同的职业,顶点之间的有向边表示两个人之间交接任务的方向,每条边上都有一个时间区间 (s,f) 表示任

收稿日期:2022-08-18　　**修回日期**:2022-12-20
作者简介:金浩宇(1995-),男,硕士研究生,研究方向为知识图谱、大数据;霍　宏(1972-),女,讲师,博士,研究方向为知识图谱、遥感图像处理、图像理解与分析;通讯作者:方　涛(1964-),男,教授,博士,研究方向为遥感图像理解、深度学习与知识图谱。

务交接的开始时间 s 与结束时间 f。



(a)任务合作时序图　　(b)任务合作模式图 Q_T　　(c)图模式匹配结果

图 1　任务合作时序图

假设有一个任务合作模式图 Q_T,如图 1(b)所示,Q_T 中每条边 e_i 都具有一个时序优先级,时序优先级规定了低优先级边的开始时间一定要晚于高优先级边的结束时间,即模式图 Q_T 隐含着一种时序约束,该文称其为时序模式图,其中 A 先与 B 交接任务,然后 A 在接收了 B 和 E 的交接信息后,再与 C 进行交接,最后 C 与 D 完成任务交接。现要在如图 1(a)所示的任务合作时序图上查询与 Q_T 相同或相似的子图。如果采用静态的图模拟匹配方法,可以得到的一个匹配结果的边集合为 $\{e_1,e_2,e_3,e_4,e_6,e_7\}$,显然,由于没有考虑边具有的时序属性,边 e_7 并不满足模式图 Q_T 中的时序约束,而正确的匹配结果如图 1(c)所示。除了时序约束外,观察图 1(b),不难发现模式图 Q_T 中还存在环路,如 $A \rightarrow B \rightarrow A$ 等,而现有的时态图模式匹配方法无法处理模式图中的环路。因此,需要研究适合于时态图并能处理环路的时序模式图匹配方法。

针对时序图特有的时序属性兼顾其拓扑结构,提出了时序优先级约束的时序模式图强模拟匹配算法,该算法在匹配过程中,不仅对拓扑结构进行检查,还将各条边的时序优先级作为局部约束,并设置了冗余顶点过滤规则来缩小搜索范围,以及对时序检查的队列顺序进行优化使得算法能够提前剪枝,减少了计算复杂度,所提出的方法兼具子图同构和图模拟的优势,实现了既准确又高效的图匹配。在三个时序数据集上的大量实验表明,相比传统的强模拟算法,所提算法能够有效过滤错误结果,并且在不同规模的数据图上均具有良好的性能表现。

1　相关工作

1.1　近似匹配

近似匹配放松了图匹配条件,成为广泛采用的一种图模式匹配方法,主要包括图模拟[5]、双向模拟[6]、强模拟[7]、受限模拟[8]、量化模拟[9-10]等。图模拟要求匹配每个节点以及其子节点。双向模拟在图模拟的基础上还要匹配父节点。强模拟在双向模拟基础上还需考虑匹配子图的局部特性。受限模拟是指在模式图上对跳跃次数进行了限制。量化模拟关注了图上的具有计数量词的表达模式。上述的近似匹配方法都只考虑了图的拓扑结构,而忽略了图上的时间信息,无法用来解决在时态图上的匹配问题。

1.2　时态图及其图模式匹配

时态图主要用于对实体及其之间涉及的时态关系进行建模,是在静态图的基础上演变的一种新型图数据[11]。目前,关于时态图的研究主要包括可达性查询[12]、最小生成树[13]、最短时态路径[14]等。

时态图上的图模式匹配问题的研究才刚刚起步,Xu 等[15]研究了时态图同构匹配,提出了时间约束的模式图匹配算法(Time – Constrained Graph Pattern Matching,TCGPM),但同构匹配要求匹配到与查询模式图完全相同的子图,过于严格。Song 等[16]研究了图流上的事件模式匹配,将时间窗口概念引入了子图匹配并提出了 DDST(Degree – Preserving Dual Simulation with Timing Constraints)算法,且基于标签过滤与度约束的思想提出了两个优化算法 DDST – SIGNATURE 和 COLORING 以提升算法的运行效率,但是这些算法将给定时间窗口的时态图快照作为静态图处理,容易导致遗漏部分结果。Ma 等[17]根据受限模拟和时态路径定义了时态图上的图模式匹配并基于连接的方式枚举匹配结果,但是这种方法的匹配条件过于宽松,并且不允许模式图中出现环路。

2　模式图匹配与时序模式图匹配的形式化描述

双向模拟和强模拟是经典的模式图匹配算法,两者的形式化描述如下:

定义 1　双向模拟。给定一个模式图 $Q=(V_q,E_q,L_q)$ 和一个数据图 $G=(V,E,L)$,如果存在一个二元匹配关系 $S \subseteq V_q \times V$ 且满足下列条件,则说 G 通过双向模拟匹配 Q:

(1)每一对 $(u,v) \in S$,u 和 v 都有相同的标签;

(2)对每个 $v \in V_q$,存在 $u \in V$ 满足 $(u,v) \in S$ 且

每一条边 $e'(v,v_1) \in E_q$ 都存在一条边 $e(u,u_1) \in E$ 使得 $(u_1,v_1) \in S$；

(3)对每个 $v \in V_q$，存在 $u \in V$ 满足 $(u,v) \in S$ 且每一条边 $e'(v_2,v) \in E_q$ 都存在一条边 $e(u_2,u) \in E$ 使得 $(u_2,v_2) \in S$。

定义 2 强模拟。给定一个模式图 $Q=(V_q,E_q,L_q)$ 和一个数据图 $G=(V,E,L)$，若在 G 中存在一个顶点 v 和一个连通子图 $G_s(V_s,E_s)$ 满足以下条件，则说 G 通过强模拟匹配 Q：

(1) Q 对 G_s 满足双向模拟匹配，且有最大二元匹配关系 S，即 Q 在 G 上的任意二元匹配关系 S_A 都有 $S_A \subseteq S$。

(2) G_s 是 S 构成的图，即(a)有一个顶点 $w \in V_s$ 当且仅当它存在于 S。(b)有一条边 $e(w,w') \in E_s$ 当且仅当在 Q 中存在一条边 $e'(u,u')$ 使得 $(u,w) \in S$ 且 $(u',w') \in S$。

(3)存在一个球 $\hat{G}[v,d_q]$，球中任意顶点 v' 到 v 的最短距离 $d \leqslant d_q$，d_q 为 Q 的直径，且顶点之间的边与 G 中对应顶点之间的边相同，$\hat{G}[v,d_q]$ 将 G_s 包含其中。

时态图的边具有时间属性，其形式化描述为：

定义 3 时态图。给定一个有向带标签的图 $G_T=(V,E,L)$，其中 V 是顶点的集合；$E \in V \times V$ 是有向时态边的集合；L 是一个标签函数，可将每个顶点 $u \in V$ 映射到其对应的标签 $L(u)$。假设图中一条有向时态边 $e_i \in E$ 连接了两个顶点 $u,v \in V$，可表示为一个四元组 (u,v,s_i,f_i)，该四元组的含义为两个顶点 u 和 v 的关系从时间点 s_i 开始到时间点 f_i 结束。

时态图的模式图及其强模拟匹配算法的形式化描述如下：

定义 4 时序模式图。一个时序模式图表示为 $Q_T=(V_q,E_q,L_q)$，其中 V_q 是顶点的集合，L_q 是一个标签函数，$E_q \in V_q \times V_q$ 是有向时序优先级边的集合。一条有向时序优先级边 $e_i \in E_q$ 可表示为一个三元组 (u,v,p_i)，其中 p_i 为 e_i 的时序优先级。时序模式图不关心具体的时间范围，而是通过时序优先级规定每条边代表的事件发生的先后顺序，时序优先级低的事件必须在时序优先级高的事件结束之后发生。

定义 5 时序强模拟匹配。给定一个时序模式图 $Q_T=(V_q,E_q,L_q)$ 和一个时态数据图 $G_T=(V,E,L)$，如果满足以下条件，则称 G 通过时序强模拟匹配 Q_T：

(1)存在一个顶点 $v \in V$，以 v 为球心，Q_T 的直径 d_q 为半径的球 $\hat{G}[v,d_q]$ 包含的子图 G_s 与 Q_T 满足双向模拟匹配。

(2)对任意两条边 $e_i'(u_i,v_i,p_i)$，$e_j'(u_j,v_j,p_j) \in E_q$，若 $p_i < p_j$，则 G_s 中存在一对对应的 $e_i,e_j \in E$，使得 $f_i < s_j$。

3 文中算法

该文提出的时序优先级约束的时序模式图强模拟匹配算法(Temporal Priority Constrained Graph Pattern Strong Simulation Matching，TPC-GPSSM)针对时序图特有的时序特性，将各条边的时序优先级作为局部约束，在匹配过程中，能够同时兼顾时序图边具有的时序优先级及其拓扑结构，也适用于带环路的时序模式图的匹配。

3.1 TPC-GPSSM 算法

TPC-GPSSM 算法主要先在时态图上构建球，然后在球上进行时序优先级约束的强模拟匹配。在时态图上构建球采用广度优先遍历(Breath First Search，BFS)算法[18]，以每个顶点为球心，构造半径为时序模式图直径的球，可将构建好的每个球看作是时态图的一个子图。然后，将每个球与时序模式图进行双向模拟匹配，在匹配过程中，一方面要计算每个球其所包含的二元匹配关系，并检查时序优先级约束，找到最大的二元匹配关系。另一方面，在匹配的同时还进行拓扑结构检查，以保证匹配的子图与时序模式图匹配。

算法 1 给出了 TPC-GPSSM 算法的伪代码。首先，TPC-GPSSM 使用 BFS 算法[18]以时态图的每个顶点 v 为球心，模式图的直径为半径，建立球 $\hat{G}$(见第 2-3 行)，并放入球集合 B 中。接着，对 B 中的每一个球计算它的最大二元匹配关系，与时序模式图进行双向模拟匹配(DualSimMatch)得到候选顶点集合 $\mathrm{sim}(v)$ 和候选边集合 $\mathrm{sim}(e')$，再进行时序优先级局部约束(TemporalFilter)，移除不满足匹配约束条件的顶点和边，并进一步进行拓扑结构检查(CheckTopology)，保证在移除某些时态边后球中的子图仍满足对时序模式图的匹配，并根据候选集构造子图 G_s(见第 4-7 行)。若此时子图 G_s 为非空，则作为一个匹配结果放入匹配结果集合 S 中，当所有子图遍历结束后，即得到所有满足时序约束的匹配结果集合 S(第 8-10 行)。

算法 1 TPC-GPSSM

输入：时序模式图 $Q_T=(V_q,E_q,L_q)$ 及其直径 d_q，时态图 $G_T=(V,E,L)$

输出：匹配结果集合 S

1. $S:=\varnothing$, $B:=\varnothing$

2. for each v in V do

3. $B:=B \cup \hat{G}[v,d_q]$

4. for each $\hat{G}[v,d_q]$ in B do

5. $\mathrm{sim}(v)$, $\mathrm{sim}(e'):=$ DualSimMatch(Q_T, $\hat{G}[v,d_q]$)

6. $\mathrm{sim}(v)$, $\mathrm{sim}(e'):=$ TemporalFilter (Q_T, $\mathrm{sim}(v)$, $\mathrm{sim}(e')$)

7. G_s: = CheckTopology(Q_T , sim(v) , sim(e'))
8. if $G_s \neq \varnothing$ then
9. S: = $S \cup G_s$
10. return S

算法 2 给出了球上进行双向模拟的伪代码 DualSimMatch,主要完成将球 $\hat{G}[v,d_q]$ 中包含的子图与 Q_T 进行双向模拟匹配,其输入是时序模式图 Q_T 以及以 v 为中心, d_q 为半径的球 $\hat{G}[v,d_q]$ 。首先,计算 V_q 中每一个顶点对应的候选顶点集合 sim(v) 。如果 u 与 v 有相同的标签,即 $L_q(u) = L(v)$,且包含于球中,则 $u \in$ sim(v) (见第 1-2 行)。同理,对 E_q 中的每一条边 e' 计算其对应的候选边集合 sim(e') (见第 3-4 行)。然后逐一检查每个候选顶点并移除不符合条件的顶点以及该顶点构成的边(见第 5-12 行)。一个顶点 $u \in$ sim(v) 必须同时满足下列条件:(1)如果 v 有一个后继节点 v' ,则 u 必须有后继节点 $u' \in$ sim(v') ;(2)如果 v 有一个前驱节点 v'' ,则 u 必须有前驱节点 $u'' \in$ sim(v'') 。否则将 u 从 sim(v) 中移除,并将 u 构成的边从 sim(e') 中移除。最后,得到候选顶点集合 sim(v) 和候选边集合 sim(e') 。

算法 2 DualSimMatch

输入:时序模式图 Q_T ,球 $\hat{G}[v,d_q]$
输出:候选顶点集合 sim(v) ,候选边集合 sim(e')
1. for each $v \in V_q$ do
2. sim(v) : = { u | u is in $G[v,d_q]$ and $L_q(u)$ }
3. for each $e(v,v_1) \in E_q$ do
4. sim(e') : = { e | $e(u,u_1)$ is in $G[v,d_q]$, $u \in$ sim(v) , $u_1 \in$ sim(v_1) }
5. while there are changes do
6. for each $u \in$ sim(v) do
7. for each $e(v,v')$ in E_q do
8. if $\nexists e(u,u') \in$ sim(e') with $u' \in$ sim(v') then
9. Remove u and edges containing u from sim(v) and sim(e')
10. for each $e(v',v)$ in E_q do
11. if $\nexists e(u',u) \in$ sim(e') with $u' \in$ sim(v') then
12. Remove u and edges containing u from sim(v) and sim(e')
13. return sim(v) , sim(e')

算法 3 给出了时序优先级局部约束的伪代码 TemporalFilter,主要功能是对所有时态边进行检查,根据模式图的时序要求移除不符合条件的边。逐一检查候选顶点集合和候选边集合是否满足时序优先级约束,将不满足约束的顶点和边删除,直到没有顶点和边再被删除为止(见第 1-8 行)。一条边 $e(u,u') \in$ sim(e') 必须满足下列条件,否则会被从 sim(e') 中删除(见第 2-5 行):(1) u , u' 都在 sim(v) 中;(2)若 e'_1 的时序优先级小于相邻边 e'_2 的时序优先级,则必须存在 $e_2 \in$ sim(e'_2) ,使得 $s_1 > f_2$。若 sim(e') 为空,或构成边的一个顶点被删除了,则删除另一个顶点以及其关联的边(见第 6-7 行)。最后,将更新的 sim(v) 和 sim(e') 返回。

算法 3 TemporalFilter

输入:时序模式图 Q_T ,候选顶点集合 sim(v) ,候选边集合 sim(e')
输出:候选顶点集合 sim(v) ,候选边集合 sim(e')
1. while there are changes do
2. for each $e'_1 = (v,v_1,p_1) \in E_q$ do
3. for each $e_1 = (u,u_1,s_1,f_1) \in$ sim(e'_1) do
4. if $p_1 > p_2$ and $s_1 < f_2$ then
5. sim(e'_1) : = sim(e'_1) \ { e_1 }
6. if sim(e'_1) = $\varnothing$ or $\nexists$ { e_1 | $e_1 \in$ sim(e'_1), $u \in$ sim(v) , $u_1 \in$ sim(v_1) } then
7. Remove u , u_1 and all edges containing u , u_1 from sim(v) and sim(e')
8. return sim(v) , sim(e')

算法 4 给出了拓扑结构检查的伪代码 CheckTopology,主要功能是对球内的子图进行检查,去除在时序优先级约束下因删除时态边而产生的不符合 Q_T 拓扑结构要求的顶点和边。对顶点候选集 sim(v) 中的任意顶点 u ,首先检查其后继关系,后继顶点集合为 succ(u) ,后继顶点候选集为 sim(v') ,若 succ(u) $\cap$ sim(v') = $\varnothing$,则将该节点及其关联的边从候选集合中删除;然后检查其前驱关系,前驱顶点集合为 pred(u) ,前驱顶点候选集为 sim(v'') ,若 pred(u) $\cap$ sim(v'') = $\varnothing$,则将相关顶点和边删除(见第 1-5 行)。若顶点和边候选集均非空,则将其构造成一个子图 G_s (见第 6-8 行)。

算法 4 CheckTopology

输入:时序模式图 Q_T ,候选顶点集合 sim(v) ,候选边集合 sim(e')
输出:匹配子图 G_s
1. While there are changes do
2. for each $u \in$ sim(v) do
3. if succ(u) $\cap$ sim(v') = $\varnothing$ or pred(u) $\cap$ sim(v'') = $\varnothing$ then
4. sim(v) : = sim(v) \ { u }
5. remove all edges containing u from sim(e')
6. if sim(v) , sim(e') $\neq \varnothing$ then
7. construct G_s from sim(v) , sim(e')
8. return G_s

TPC-GPSSM 在球的建立过程花费了很多时间,主要因为每个球包含了很多无用的拓扑结构信息,并且在时序检查过程中存在重复检查的问题。如果能减少每个球的计算量, 避免时序的重复检查,将大大提升算法的效率。

3.2　TPC-GPSSM 算法优化方法

针对上文提及的 TPC-GPSSM 算法存在很多重复

计算的问题,进一步对球的建立过程、时态优先级约束等方面进行优化,算法 5 给出了优化后的 TPC-GPSSM 算法的伪代码,主要包括:

第一,在球的建立过程中,TPC-GPSSM 算法对数据图中每个顶点,都作为中心建立了球,然而其中大量的球并没有包含能够匹配模式图的子图,对这些球的计算浪费了大量时间。在建球之前先进行双向模拟匹配(见第 2 行),能够过滤掉大量的节点与边,这样既能减少需要建立的球的数量,也能使每个球包含的子图规模减小。

第二,在对球心的选择方面,只选择与模式图中距离最远的两个节点相匹配的节点(见第 3-8 行),如图 1(b)所示,E、D 是模式图中距离最远的两个点,以 E 对应的顶点 E_2 为球心建球得到的结果与以 A_1 为球心建立球得到的结果相同,因此计算以 A_1 为球心的球是冗余的。

第三,在时序优先级约束过程中,先对模式图中的边根据时序优先级进行排序,从高优先级的边开始依次检查,每一条边只检查与其相邻的时序优先级更高的边的之间的时序关系,可避免重复时序检查。

算法 5 优化后的 TPC-GPSSM 算法

输入:时序模式图 $Q_T=(V_q,E_q,L_q)$ 及其直径 d_q,时态图 $G_T=(V,E,L)$

输出:子图集合 S

1. $S:=\varnothing$, $B:=\varnothing$

2. $\mathrm{sim}(v),\mathrm{sim}(e'):=\mathrm{DualSimMatch}(Q_T,\hat{G}[v,d_q])$

3. v_{f1}, v_{f2} are the farthest two nodes in Q

4. for each $v\in\mathrm{sim}(v_{f1})$ or $\mathrm{sim}(v_{f2})$ do

5. build $\hat{G}[v,d_q]$ with $u\in\mathrm{sim}(v)$ and $e\in\mathrm{sim}(e')$

6. $B:=B\cup\{\hat{G}[v,d_q]\}$

7. for each $\hat{G}[v,d_q]$ in B do

8. $\mathrm{sim}b(v):=\{u\mid u\in\mathrm{sim}(v),u\text{ in }\hat{G}[v,d_q]\}$

9. $\mathrm{sim}b(e'):=\{e\mid e\in\mathrm{sim}(e'),e\text{ in }\hat{G}[v,d_q]\}$

10. for p from the highest priority to the lowest priority do

11. $\mathrm{sim}b(v),\mathrm{sim}b(e'):=\mathrm{TemporalFilter}(Q_T,\mathrm{sim}b(v),\mathrm{sim}b(e))$

12. $G_s:=\mathrm{CheckTopology}(Q_T,\mathrm{sim}b(v),\mathrm{sim}b(e))$

13. if $G_s\neq\varnothing$ then

14. $S:=S\cup G_s$

15. Return S

优化后的 TPC-GPSSM 算法大幅减少了球的数量,同时避免了球之间因为交集而进行的重复计算,且保证了对每条时序边只检查一次,解决了重复检查的问题。

4 实验与结果分析

本章节对分别从算法有效性以及算法效率两个维度来评估提出的算法,并对普通的强模拟匹配算法[19](General Strong Simulation,GSS)与 TPC-GPSSM 及优化后的 TPC-GPSSM 算法进了比较。实验采用三个时态数据集,分别是:(1)BMC[20],一个小学师生相互接触时态网;(2)Enron[21],一个电子邮件传输网络;(3)Wikitalk-Russian[22],一个 Russian Wikipedia 上的交流网络。表 1 列出了这些数据集的详细参数,其中 $|V|$、$|E|$、$|T|$、#labels 分别是数据集的顶点数量、边的数量、边的平均时间区间、标签数量。

表 1 数据集

数据集	$\lvert V\rvert$	$\lvert E\rvert$	$\lvert T\rvert$	#labels
BMC	242	251 546	20	11
Enron	87 273	1 148 072	20	26
Wikitalk-Russian	457 017	2 282 055	35	26

模式图由一个自行开发的模式图生成器提供。给定模式图的顶点数量和标签集合,模式图生成器随机生成包含不同时序优先级数量的模式图。实验对每个数据集分别生成了 100 个模式图,对其平均结果进行评估。

4.1 算法有效性实验

首先,对 TPC-GPSSM 算法的有效性进行了实验验证。TPC-GPSSM 算法相比传统强模拟匹配算法,能够有效地过滤不符合时态要求的边。为了定量地描述算法有效性,定义了一个评价指标——时态边聚合度(Temporal Edge Closeness,TEC)。

给定一个时序模式图 $Q_T=(V_q,E_q,L_q)$,以及一组作为匹配结果返回的子图 $\{M_1,M_2,\cdots,M_n\}$,TEC 定义如下式:

$$\mathrm{TEC}=\sum_{j=1}^{n}|E_{M_j}|/|E_q|$$

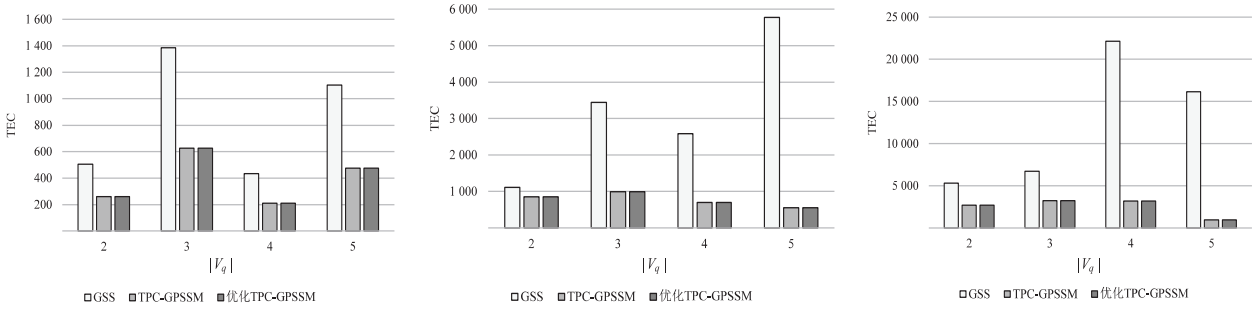
其中,$|E_{M_j}|$ 是 M_j 中时态边的数量,$|E_q|$ 是模式图的边数量,TEC 的值越小说明对时态边的过滤效果越明显。对三个数据集的实验结果如图 2 所示,实验改变时序模式图顶点数量 $|V_q|$ 从 2 到 5,分别在三个数据集上进行测试。其中,GSS 由于没有考虑时态边的顺序,所以得到了最大的 TEC 值,而 TPC-GPSSM 及其改进算法对不合模式图需求的时态边进行了过滤,得到了更优的 TEC 值。

4.2 算法效率实验

这里通过二个实验,分别考察时序模式图、不同时序优先级数量对 TPC-GPSSM 算法效率和性能的影响。

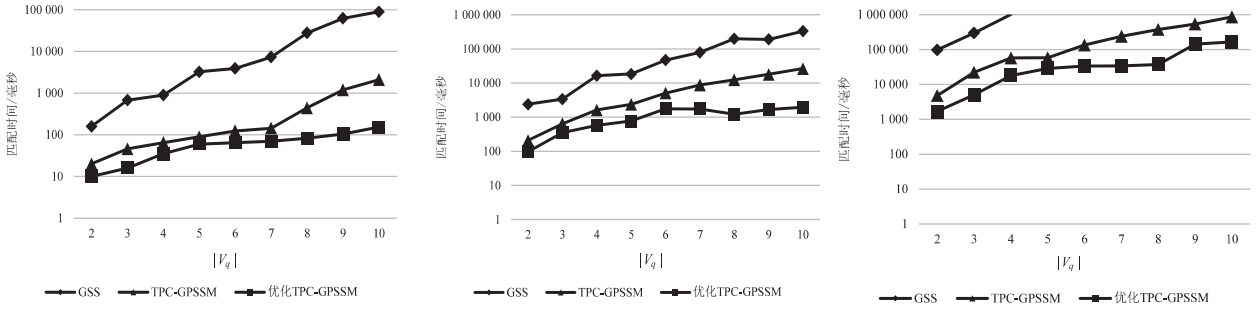
(1)时序模式图顶点数量 $|V_q|$ 对算法效率的影响。设置了 2 到 10 不同大小的 $|V_q|$,在三个数据集上进行了实验。实验结果如图 3 所示,可以观察到,三

种算法的匹配时间都随着 $|V_q|$ 的增大而增加,这是因为当 $|V_q|$ 增大时,需要匹配的节点和边的数量也同时增加,这将增加计算的时间花销。优化后的 TPC-GPSSM 算法采用了先进行双向模拟匹配,再建立球进行局部性检查,使得球的数量大幅减少,取得了更加良好的性能表现。



(a)BMC　(b)Enron　(c)Wikitalk-Russian

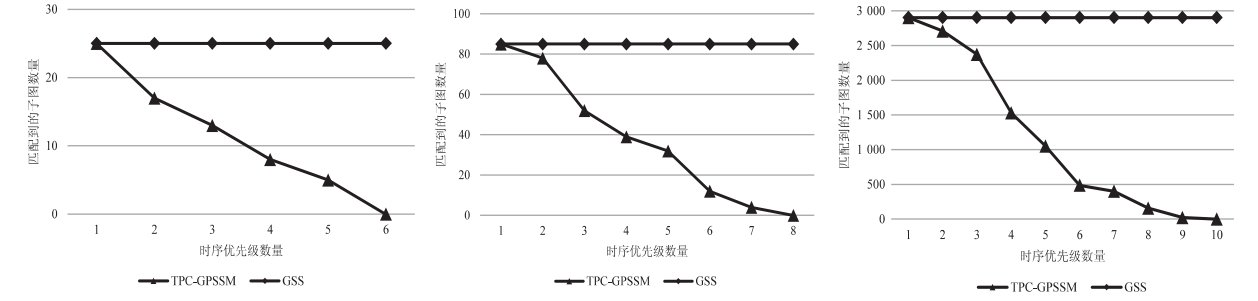
图 2　不同数据集上匹配质量的评价



(a)BMC　(b)Enron　(c)Wikitalk-Russian

图 3　时序模式图大小对算法效率的影响

(2)不同时序优先级数量对匹配结果的影响。实验中固定模式图的 $|V_q|$ 为 10,改变其中的时序优先级数量,实验结果如图 4 所示。当时序优先级数量增加时,意味着匹配条件更加严格,符合条件的子图数量也随之减少,而 GSS 不考虑时序关系,时序优先级数量不影响它的匹配结果。



(a)BMC　(b)Enron　(c)Wikitalk-Russian

图 4　时序优先级数量对匹配结果的影响

5　结束语

针对时态图上的模式图匹配,提出了一种时序优先级约束的时序模式图强模拟匹配算法(TPC-GPSSM),该算法将时序优先级作为局部约束,在模式图的匹配过程中考虑了时态图中不同时态边之间的时序优先级,并兼顾图的拓扑结构,通过设置冗余顶点过滤规则来缩小搜索范围,优化时序检查的队列顺序,以达到提前剪枝、减少计算复杂度的目的。提出并利用时态边聚合度作为评价指标,在三个时序数据集上的大量实验表明,相比普通的强模拟算法,所提算法能够有效过滤错误结果,并且在不同规模的数据图上均具有良好的性能表现。该算法可应用于涉及前后时序关系的场景,如任务交接分析、交通网络分析等,特别是在任务交接中存在环路的情况下该算法也能进行有效的处理。在下一步的工作中,一个方向是考虑不同时间窗口的约束,另一个方向是在分布式环境中开发更性能更高的算法。

参考文献:

[1] LEI Z, LEI C, ZSU M T. Distancejoin: pattern match query in a large graph database[J]. Proceedings of the VLDB En-

dowment,2009,2(1):886-897.
[2] WANG X,CHAI L,XU Q,et al. Efficient subgraph matching on large rdf graphs using mapreduce[J]. Data Science and Engineering,2019,4(1):24-43.
[3] HOLME P,SARAMÄKI J. Temporal networks[J]. Physics Reports:A Review Section of Physics Letters,2012,519(3):97-125.
[4] KOSTAKOS V. Temporal graphs[J]. Physica A Statistical Mechanics & Its Applications,2009,388(6):1007-1023.
[5] HENZINGER M R,HENZINGER T A,KOPKE P W. Computing simulations on finite and infinite graphs[C]//Proceedings of IEEE 36th annual foundations of computer science. Milwaukee:IEEE,1995:453-462.
[6] MA S,CAO Y,FAN W,et al. Capturing topology in graph pattern matching[J]. Proceedings of the VLDB Endowment,2011,5(4):310-321.
[7] MA S,CAO Y,FAN W,et al. Strong simulation:capturing topology in graph pattern matching[J]. ACM Transactions on Database Systems,2014,39(1):4.
[8] FAN W, LI J, SHUAI M, et al. Graph pattern matching:from intractable to polynomial time[J]. Proceedings of the Vldb Endowment,2010,3(1):264-275.
[9] MAHFOUD H. Graph pattern matching with counting quantifiers and label-repetition constraints[J]. Cluster Computing,2020,23(3):1529-1553.
[10] SHAMIR R,TSUR D. Faster subtree isomorphism[J]. Journal of Algorithms,1999,33(2):267-280.
[11] HOLME P. Modern temporal network theory:a colloquium[J]. Physics of Condensed Matter,2015,88(9):1-30.
[12] BRITO L F A,ALBERTINI M,CASTEIGTS A,et al. A dynamic data structure for temporal reachability with unsorted contact insertions[J]. Social Network Analysis and Mining,2022,12(1):22.
[13] KHANNA G,SOH S,CHATURVEDI S K,et al. On enumeration of spanning arborescences and reliability for network broadcast in fixed-schedule dynamic networks[J]. IEEE Transactions on Network Science and Engineering,2021,7(4):2980-2996.
[14] 张天明,徐一恒,蔡鑫伟,等. 时态图最短路径查询方法[J]. 计算机研究与发展,2022,59(2):362-375.
[15] XU Y,HUANG J,LIU A,et al. Time-constrained graph pattern matching in a large temporal graph[C]//Joint international conference on APWeb-WAIM. Beijing:Springer,2017:100-115.
[16] SONG C,GE T,CHEN C,et al. Event pattern matching over graph streams[J]. Proceedings of the VLDB Endowment,2014,8(4):413-424.
[17] MA Y,YUAN Y ,LIU M,et al. Graph simulation on large scale temporal graphs[J]. GeoInformatica,2020,24(1):1-22.
[18] DIESTEL R. Graph theory[J]. Mathematical Gazette,2000,173(502):67-128.
[19] 沈嘉思. 强模拟在带权有向图的扩展及其匹配结果的排序[J]. 现代电信科技,2012(9):53-57.
[20] GEMMETTO V,BARRAT A,CATTUTO C. Mitigation of infectious disease at school:targeted class closure vs school closure[J]. Bmc Infectious Diseases,2014,14(1):695.
[21] KLIMT B B,YANG Y. The Enron corpus:a new dataset for email classification research[C]//European conference on machine learning. Pisa:KDNet,2004:217-226.
[22] SUN J,KUNEGIS J,STAAB S. Predicting user roles in social networks using transfer learning with feature transformation[C]//International conference on data mining. Barcelona:IEEE,2016:128-135.