(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 106100921 A
(43)申请公布日 2016.11.09

(21)申请号 201610406969.8

(22)申请日 2016.06.08

(71)申请人 华中科技大学
地址 430074 湖北省武汉市洪山区珞喻路1037号

(72)发明人 施展 冯丹 王子毅 余静 彭亚妹 于瑞丽

(74)专利代理机构 华中科技大学专利中心 42201
代理人 李智

(51)Int.Cl.
H04L 12/26(2006.01)
H04L 12/24(2006.01)

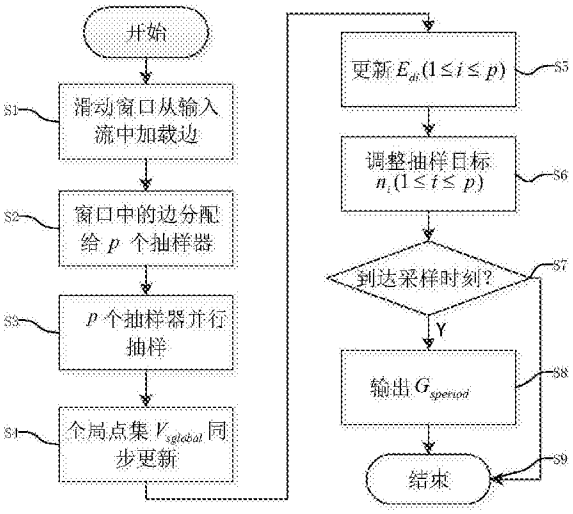
权利要求书2页 说明书6页 附图2页

(54)发明名称

基于点信息同步的动态流式图并行抽样方法

(57)摘要

本发明提供一种基于点信息同步的动态流式图并行抽样方法，具体为：S1.流式边到达滑动窗口，判断窗口是否满，如果不满则执行S1，否则执行S2；S2.将滑动窗口中的边按序随机分配给多个抽样器；S3.多个抽样器对被分配的边并行处理得到抽样子图；S4.读取抽样器的点集，去除重复的点，刷新全局点集合；S5.更新全局点推导的边集；S6.调整抽样目标点集数量；S7.如果未到设定的采集抽样子图时间点，更新滑动窗口，返回S1；否则执行S8；S8.根据每个抽样器的抽样结果合成抽样子图。本发明在快速获得抽样子图的同时，保证抽样子图与原图的特性相似度高，解决了动态流式图串行抽样算法处理时间长、不能满足实时性要求的问题。



CN 106100921 A

1.一种基于点信息同步的动态流式图并行抽样方法，其特征在于，包括以下步骤：

S1.流式边e＝(u,v)到达滑动窗口，判断滑动窗口是否满，如果不满则执行S1，否则执行S2；

S2.将滑动窗口中的边按序随机分配给p个抽样器，直至分配完毕；

S3.p个抽样器对被分配的边并行处理，得到$G_{sk}=(V_{sk},E_{sk},E_{dk})$；第k个抽样器抽的抽样子图为$G_{sk}=(V_{sk},E_{sk},E_{dk})$，1≤k≤p，其中，$V_{sk}=\{v_{sk1},v_{sk2},\cdots,v_{skn'}\}$为子图的点集合，$v_{ski}$，1≤i≤n′为抽样子图中的点，且抽样点集的大小$|V_{sk}|=n'=n/p$，其中n为抽样目标点数量；$E_{sk}=\{e_{sk1},e_{sk2},\cdots,e_{skm}\}$为子图的边集合，其中的$e_{ski}$，1≤i≤m为抽样子图中的边；$E_{dk}=\{e_{dk1},e_{dk2},\cdots,e_{dkt}\}$为子图的推导边集合，其中$e_{dki}$，1≤i≤t为抽样子图中的依据全局点集$V_{sglobal}$得到的推导边；

S4.全局点信息同步：依次读取每个抽样器的点集V_{sk}，去除重复的点，刷新全局点集合$V_{sglobal}$；

S5.更新全局点推导的边集：每个抽样器利用更新后的全局点信息$V_{sglobal}$，对于$\exists e \in E_{dk}$，1≤k≤p，如果e的端点不在$V_{sglobal}$中，则删除e；

S6.调整抽样目标点集数量：当$|V_{sglobal}|<n$，那么均等地增加抽样器的抽样目标点数量n；如果$|V_{sglobal}|>n$，那么均等地减小抽样目标点数量n；

S7.如果未到设定的采集抽样子图时间点，按照抽样目标点数量n增大则增大滑动窗口，目标点数量n减小则减小滑动窗口的原则更新滑动窗口大小，返回步骤S1；否则执行步骤S8；

S8.根据每个抽样器的抽样结果合成抽样子图：抽样子图表示为：$G_{speriod}=(V_{speriod},E_{speriod})$，其中，$V_{speriod}=V_{s1}\cup V_{s2}\cup\cdots\cup V_{sp}$为所有抽样器中点集的并集，$E_{speriod}=E_{s1}\cup E_{s2}\cup\cdots\cup E_{sp}\cup E_{d1}\cup E_{d2}\cup\cdots\cup E_{dp}$为所有抽样器中边集和全局推导边集的并集；

S9.结束。

2.根据权利要求1所述的基于点信息同步的动态流式图并行抽样方法，其特征在于，所述步骤S3中每个抽样器进行并行处理的步骤如下：

a)流式边e＝(u,v)到达某抽样器，该抽样器判定是否会产生点替换，如果发生点替换，则执行b)，否则执行f)；

判定原则为：

i.若流式边中的点$u\in V_{sk}\cup V_{sglobal}$，$v\in V_{sk}\cup V_{sglobal}$，不会引起$V_{sk}$添加新点，不发生替换；

ii.若流式边中的点$u\in V_{sk}\cup V_{sglobal}$，$v\notin V_{sk}\cup V_{sglobal}$或$u\notin V_{sk}\cup V_{sglobal}$，$v\in V_{sk}\cup V_{sglobal}$，且现有点的个数$|V_{sk}|<n'$，则不发生替换；否则，点v或者u需要添加到$V_{sk}$中并且替换掉一个现有的点；

iii.若流式边中的点$u\notin V_{sk}\cup V_{sglobal}$，$v\notin V_{sk}\cup V_{sglobal}$，现有点的个数$|V_{sk}|<n'-1$，则不发生替换；否则，u和v都需要添加到$V_{sk}$中并替换掉两个现有的点；

b)每个抽样器都各自独立根据该抽样器内抽样子图G_{sk}，1≤j≤p，中的点的度特性，确定点替换概率函数$f_k(d_i)$，$d_i\in D_j$，其中D_j是第j个抽样器中所有节点的度的集合；根据该替换概率函数计算点v_i被替换的概率$p_{v_i}=f_k(d_{v_i})$，得到替换概率集合$P=\{p_{v_1},p_{v_2},\cdots,p_{v_{n'}}\}$，

其中d_{v_i}是点v_i的度，且有$\sum_1^n p_{v_i}=1$；其中$D_k=\{d_{k1},d_{k2},\cdots,d_{kn'}\}$为点集合中点的度分布；其中要求替换概率函数$f_k(d_i)$在作用域$[1,d_{max}]$内单调递减，其中$d_{max}$为度分布集合中最高的度数；

c)每个抽样器均采用遗传算法中的选择算法select(P)，其中P为b)中计算得到的点替换概率集合，选取被替换的点r；

d)每个抽样器根据替换原则，判断c)选择的点r是否符合被替换要求，若符合则转至e)；否则转至c)；

替换原则如下：

i.上述a)的ⅱ)情况下，在选择被替换点时，不能选择新增边中的点，并且在后续的孤立点删除时也不能删除新增边中的点；

ii.上述a)的ⅲ)情况下，假定先添加u，再添加v；先添加u时，V_{sk}中没有与其相关联的点，替换并没有限制；再添加v时，由于V_{sk}中存在与其相连的点u，被替换的点不能为u；

e)每个抽样器从V_{sk}中删除r，并从E_{sk}和E_{ak}中删除与r相关联的边；再从V_{sk}中删除孤立点；孤立点删除要求不能删除新增点中的第一个点；

f)每个抽样器把新增点和边加入子图G_{sk}中；其中增加点和边到子图的原则是：如果u,v均在点集V_{sk}中，那么将e=(u,v)加入边集E_{sk}中；如果u,v中一个在点集V_{sk}，另一个在全局点集合$V_{sglobal}$中，那么将e加入推导边集E_{dk}中；如果u,v均在全局点集合$V_{sglobal}$而不在点集V_{sk}中，不增加点也不增加边到子图中。

3.根据权利要求2所述的基于点信息同步的动态流式图并行抽样方法，其特征在于，替换概率函数$f_k(d_i)$为反比例函数。

4.根据权利要求2所述的基于点信息同步的动态流式图并行抽样方法，其特征在于，选择算法select(P)为遗传算法中的比例选择算法。

5.根据权利要求2所述的基于点信息同步的动态流式图并行抽样方法，其特征在于，所述步骤S6调整抽样目标点集数量的具体过程为：当$|V_{sglobal}|<n$，均等地增加抽样器的抽样目标点数量即$n=n+(n-|V_{sglobal}|)/p$；如果$|V_{sglobal}|>n$，均等地减小抽样目标点数量即$n=n-(|V_{sglobal}|-n)/p$。

6.根据权利要求5所述的基于点信息同步的动态流式图并行抽样方法，其特征在于，更新滑动窗口的大小为$(2n-|V_{sglobal}|)/2$。

基于点信息同步的动态流式图并行抽样方法

技术领域

[0001] 本发明属于流式图数据抽样技术领域，更具体地，涉及一种基于点信息同步的动态流式图并行抽样方法。

背景技术

[0002] 图已经成为了表达现实世界中对象和关系的一种无处不在且必不可少的数据结构，各种各样的应用包括社交网络、生物科学网络、万维网等都能建模成图。然而，现实中的很多应用程序是不断发生变化的，建模而成的图结构也会相应地发生变化，即随着时间的推移，可能会发生点、边增加操作或者删除操作。点和边的信息随着时间发生动态地改变的图就是所谓的动态图。处理动态图数据的理想方法是采用流式模型，即把原图看作是由一连串流式到达的点和边逐渐形成的。

[0003] 对动态图进行抽样使用的是流式抽样算法，现有的部分推导边抽样PIES(Partially-Induced Edge Sampling，见论文：“Space-efficient sampling from social activity streams”)算法，由于流式处理的累计迭代特性，整个抽样过程对流式边是依次进行处理的，也就是串行的，那么可想而知，抽样过程需要消耗大量的时间。对于实时性要求比较高的场景，如实时控制网络拥塞，这些串行的抽样算法远远不能满足要求，提高抽样算法执行速度非常有必要，故而需要更加快速的抽样算法。

发明内容

[0004] 针对现有技术的缺陷和迫切需求，本发明提供了一种基于点信息同步的动态流式图并行抽样方法，其目的在于，在快速获得抽样子图的同时，保证抽样子图与原图的特性相似度高，解决了动态流式图串行抽样算法处理时间长，不能满足实时性要求高的应用的问题。

[0005] 一种基于点信息同步的动态流式图并行抽样方法，包括以下步骤：

[0006] S1.流式边e＝(u,v)到达滑动窗口，判断滑动窗口是否满，如果不满则执行S1，否则执行S2；

[0007] S2.将滑动窗口中的边按序随机分配给p个抽样器，直至分配完毕；

[0008] S3.p个抽样器对被分配的边并行处理，得到$G_{sk}=(V_{sk},E_{sk},E_{dk})$；第k个抽样器抽的抽样子图为$G_{sk}=(V_{sk},E_{sk},E_{dk})$，$1\leqslant k\leqslant p$，其中，$V_{sk}=\{v_{sk1},v_{sk2},\cdots,v_{skn,}\}$为子图的点集合，$v_{ski}$，$1\leqslant i\leqslant n'$为抽样子图中的点，且抽样点集的大小$|V_{sk}|=n'=n/p$，其中n为抽样目标点数量；$E_{sk}=\{e_{sk1},e_{sk2},\cdots,e_{skm}\}$为子图的边集合，其中的$e_{ski}$，$1\leqslant i\leqslant m$为抽样子图中的边；$E_{dk}=\{e_{dk1},e_{dk2},\cdots,e_{dkt}\}$为子图的推导边集合，其中$e_{dki}$，$1\leqslant i\leqslant t$为抽样子图中的依据全局点集$V_{sglobal}$得到的推导边；

[0009] S4.全局点信息同步：依次读取每个抽样器的点集V_{sk}，去除重复的点，刷新全局点集合$V_{sglobal}$；

[0010] S5.更新全局点推导的边集：每个抽样器利用更新后的全局点信息$V_{sglobal}$，对于

$\exists e \in E_{dk}$，1≤k≤p，如果e的端点不在$V_{sglobal}$中，则删除e；

[0011] S6.调整抽样目标点集数量：当|$V_{sglobal}$|<n，那么均等地增加抽样器的抽样目标点数量n；如果|$V_{sglobal}$|>n，那么均等地减小抽样目标点数量n；

[0012] S7.如果未到设定的采集抽样子图时间点，按照抽样目标点数量n增大则增大滑动窗口，目标点数量n减小则减小滑动窗口的原则更新滑动窗口大小，返回步骤S1；否则执行步骤S8；

[0013] S8.根据每个抽样器的抽样结果合成抽样子图：抽样子图表示为：$G_{speriod}$=($V_{speriod}$,$E_{speriod}$)，其中，$V_{speriod}$=V_{s1}∪V_{s2}∪…∪V_{sp}为所有抽样器中点集的并集，$E_{speriod}$=E_{s1}∪E_{s2}∪…∪E_{sp}∪E_{d1}∪E_{d2}∪…∪E_{dp}为所有抽样器中边集和全局推导边集的并集；

[0014] S9.结束。

[0015] 进一步地，所述步骤S3中每个抽样器进行并行处理的步骤如下：

[0016] a)流式边e=(u,v)到达某抽样器，该抽样器判定是否会产生点替换，如果发生点替换，则执行b)，否则执行f)；

[0017] 判定原则为：

[0018] i.若流式边中的点u∈V_{sk}∪$V_{sglobal}$，v∈V_{sk}∪$V_{sglobal}$，不会引起V_{sk}添加新点，不发生替换；

[0019] ii.若流式边中的点u∈V_{sk}∪$V_{sglobal}$，$v \notin V_{sk} \cup V_{sglobal}$或$u \notin V_{sk} \cup V_{sglobal}$，v∈$V_{sk}$∪$V_{sglobal}$，且现有点的个数|$V_{sk}$|<n′，则不发生替换；否则，点v或者u需要添加到V_{sk}中并且替换掉一个现有的点；

[0020] iii.若流式边中的点$u \notin V_{sk} \cup V_{sglobal}$，$v \notin V_{sk} \cup V_{sglobal}$，现有点的个数|$V_{sk}$|<n′-1，则不发生替换；否则，u和v都需要添加到V_{sk}中并替换掉两个现有的点；

[0021] b)每个抽样器都各自独立根据该抽样器内抽样子图G_{sk}，1≤j≤p，中的点的度特性，确定点替换概率函数$f_k(d_i)$，$d_i \in D_j$，其中D_j是第j个抽样器中所有节点的度的集合；根据该替换概率函数计算点v_i被替换的概率$p_{v_i} = f_k(d_{v_i})$，得到替换概率集合$P = \{p_{v_1}, p_{v_2}, \cdots, p_{v_{n'}}\}$，其中$d_{v_i}$是点$v_i$的度，且有$\sum_1^n p_{v_i} = 1$；其中$D_k$={$d_{k1}$,$d_{k2}$,…,$d_{kn'}$}为点集合中点的度分布；其中要求替换概率函数$f_k(d_i)$在作用域[1,$d_{max}$]内单调递减，其中$d_{max}$为度分布集合中最高的度数；

[0022] c)每个抽样器均采用遗传算法中的选择算法select(P)，其中P为b)中计算得到的点替换概率集合，选取被替换的点r；

[0023] d)每个抽样器根据替换原则，判断c)选择的点r是否符合被替换要求，若符合则转至e)；否则转至c)；

[0024] 替换原则如下：

[0025] i.上述a)的ii情况下，在选择被替换点时，不能选择新增边中的点，并且在后续的孤立点删除时也不能删除新增边中的点；

[0026] ii.上述a)的iii情况下，假定先添加u，再添加v；先添加u时，V_{sk}中没有与其相关联的点，替换并没有限制；再添加v时，由于V_{sk}中存在与其相连的点u，被替换的点不能为u；

[0027] e)每个抽样器从V_{sk}中删除r，并从E_{sk}和E_{dk}中删除与r相关联的边；再从V_{sk}中删除孤立点；孤立点删除要求不能删除新增点中的第一个点；

[0028] f)每个抽样器把新增点和边加入子图G_{sk}中;其中增加点和边到子图的原则是:如果u,v均在点集V_{sk}中,那么将e=(u,v)加入边集E_{sk}中;如果u,v中一个在点集V_{sk},另一个在全局点集合$V_{sglobal}$中,那么将e加入推导边集E_{dk}中;如果u,v均在全局点集合$V_{sglobal}$而不在点集V_{sk}中,不增加点也不增加边到子图中。

[0029] 进一步地,替换概率函数$f_k(d_i)$为反比例函数。

[0030] 进一步地,选择算法select(P)为遗传算法中的比例选择算法。

[0031] 进一步地,所述步骤S6调整抽样目标点集数量的具体过程为:当$|V_{sglobal}|<n$,均等地增加抽样器的抽样目标点数量即$n=n+(n-|V_{sglobal}|)/p$;如果$|V_{sglobal}|>n$,均等地减小抽样目标点数量即$n=n-(|V_{sglobal}|-n)/p$。

[0032] 进一步地,更新滑动窗口的大小为$(2n-|V_{sglobal}|)/2$。

[0033] 本发明基于点信息同步的动态流式图并行抽样算法(简称为PaStS),在快速获得抽样子图的同时,保证抽样子图与原图的特性相似度高,解决了动态流式图串行抽样算法处理时间长,不能满足实时性要求高的应用的问题。与现有技术方案PIES算法相比较,对于相同规模的动态图,并行算法PaStS的执行效率相比串行PIES会提高p到p^2倍,其中,p是并行抽样器的数目。

附图说明

[0034] 图1为本发明方法流程图;

[0035] 图2为本发明实施例提供的单个抽样器处理单条边的流程示意图。

具体实施方式

[0036] 为了使本发明的目的、技术方案及优点更加清楚明白,以下结合附图及实施例,对本发明进行进一步详细说明。应当理解,此处所描述的具体实施例仅仅用以解释本发明,并不用于限定本发明。此外,下面所描述的本发明各个实施方式中所涉及到的技术特征只要彼此之间未构成冲突就可以相互组合。

[0037] 设第k($1\leqslant k\leqslant p$)个抽样器的抽样子图为$G_{sk}=(V_{sk},E_{sk},E_{dk})$,其中的$V_{sk}=\{v_{sk1},v_{sk2},\cdots,v_{skn'}\}$为子图的点集合,$v_{ski}$($1\leqslant i\leqslant n'$)为抽样子图中的点,且抽样点集的大小$|V_{sk}|=n'=n/p$,其中,n为目标抽样子图中点的个数;$E_{sk}=\{e_{sk1},e_{sk2},\cdots,e_{skm}\}$为子图的边集合,其中的$e_{ski}$($1\leqslant i\leqslant m$)为抽样子图中的边;$E_{dk}=\{e_{dk1},e_{dk2},\cdots,e_{dkt}\}$为子图的边集合,其中$e_{ski}$($1\leqslant i\leqslant t$)为抽样子图中的依据全局点集$V_{sglobal}$得到的推导边。

[0038] 如图1所示,本发明提供了一种基于点信息同步的动态流式图并行抽样方法,包括以下步骤:

[0039] S1.流式边e=(u,v)到达滑动窗口,判断滑动窗口是否满,如果不满足,则执行S1,否则执行S2;

[0040] 滑动窗口的大小与目标抽样子图中点的总数相关,初始值一般在小于目标抽样子图中点总数范围内可任意设置,后续通过循环调整逼近目标抽样子图中点的总数。

[0041] 按照一种优选的方式,初始设置滑动窗口大小为n/2,其后滑动窗口大小由上一轮抽样的S7步骤确定,n为抽样目标点数量;

[0042] S2.将窗口中的边按序随机分配给p个抽样器,直至分配完毕;

[0043] S3.p个抽样器对被分配的边进并行抽样,得到$G_{sk}=(V_{sk},E_{sk},E_{dk})$;第k个抽样器抽的抽样子图为$G_{sk}=(V_{sk},E_{sk},E_{dk})$,$1\leqslant k\leqslant p$,其中,$V_{sk}=\{v_{sk1},v_{sk2},\cdots,v_{skn'}\}$为子图的点集合,$v_{ski}$,$1\leqslant i\leqslant n'$为抽样子图中的点,且抽样点集的大小$|V_{sk}|=n'=n/p$,其中n为抽样目标点数量;$E_{sk}=\{e_{sk1},e_{sk2},\cdots,e_{skm}\}$为子图的边集合,其中的$e_{ski}$,$1\leqslant i\leqslant m$为抽样子图中的边;$E_{dk}=\{e_{dk1},e_{dk2},\cdots,e_{dkt}\}$为子图的推导边集合,其中$e_{dki}$,$1\leqslant i\leqslant t$为抽样子图中的依据全局点集$V_{sglobal}$得到的推导边。全局点集$V_{sglobak}$是指每隔程序自身定义的时间之后对所有抽样器的抽样子图中的V_{sk}($1\leqslant k\leqslant p$)的点集进行汇总之后得到的点集,推导边是指在一个抽样器k中,一个端点在抽样子图中的V_{sk}中,而另一个端点在全局点集$V_{sglobal}$中。

[0044] S4.全局点信息同步:依次读取每个抽样器的点集V_{sk}($1\leqslant k\leqslant p$),去除重复的点,刷新全局点集合$V_{sglobal}$;

[0045] S5.更新全局点推导的边集:每个抽样器利用更新后的全局点信息$V_{sglobal}$,对于$\exists e\in E_{dk}$,$1\leqslant k\leqslant p$,如果e的端点不在$V_{sglobal}$中,则删除e;

[0046] S6.调整抽样目标点集数量:当$|V_{sglobal}|<n$,那么均等地增加抽样器的抽样目标点数量,按照一种优选方式,更新$n=n+(n-|V_{sglobal}|)/p$;如果$|V_{sglobal}|>n$,那么均等地减小抽样目标点数量,按照一种优选方式,更新$n=n-(|V_{sglobal}|-n)/p$;

[0047] S7.如果未到设定的采集抽样子图时间点,那么,根据更新后的n更新滑动窗口大小,更新原则是n增大了就适度增大滑动窗口,n减小了就适度减小滑动窗口,返回步骤S1;否则执行步骤S8;

[0048] 按照一种优选方式,更新滑动窗口的大小为$(2n-|V_{sglobal}|)/2$。

[0049] S8.根据每个抽样器的抽样结果合成抽样子图:抽样子图可表示为:$G_{speriod}=(V_{speriod},E_{speriod})$,其中,$V_{speriod}=V_{s1}\cup V_{s2}\cup\cdots\cup V_{sp}$为所有抽样器中点集的并集,$E_{speriod}=E_{s1}\cup E_{s2}\cup\cdots\cup E_{sp}\cup E_{d1}\cup E_{d2}\cup\cdots\cup E_{dp}$为所有抽样器中边集和全局推导边集的并集;

[0050] S9.结束。

[0051] 所述步骤S3中每个抽样器进行并行处理的步骤如下:

[0052] a)流式边$e=(u,v)$到达某抽样器,该抽样器判定是否会产生点替换,如果会发生点替换,则执行b),否则执行f);

[0053] 判定原则为:

[0054] i.若流式边中的点$u\in V_{sk}\cup V_{sglobal}$,$v\in V_{sk}\cup V_{sglobal}$,不会引起$V_{sk}$添加新点,不发生替换;

[0055] ii.若流式边中的点$u\in V_{sk}\cup V_{sglobal}$,$v\notin V_{sk}\cup V_{sglobal}$或$u\notin V_{sk}\cup V_{sglobal}$,$v\in V_{sk}\cup V_{sglobal}$,且现有点的个数$|V_{sk}|<n'$,则不发生替换;否则,点v或者u需要添加到$V_{sk}$中并且替换掉一个现有的点;

[0056] iii.若流式边中的点$u\notin V_{sk}\cup V_{sglobal}$,$v\notin V_{sk}\cup V_{sglobal}$,现有点的个数$|V_{sk}|<n'-1$,则不发生替换;否则,u和v都需要添加到$V_{sk}$中并替换掉两个现有的点;

[0057] b)每个抽样器都各自独立根据该抽样器内抽样子图G_{sk}($1\leqslant j\leqslant p$)中的点的度特性,确定点替换概率函数$f_k(d_i)$,$d_i\in D_j$;根据该概率函数计算点v_i被替换的概率$p_{v_i}=f_k(d_{v_i})$,得到替换概率集合$P=\{p_{v_1},p_{v_2},\cdots,p_{v_n}\}$,其中$d_{v_i}$是点$v_i$的度,且有$\sum_1^n p_{v_i}=1$;其中$D_k=\{d_{k1},d_{k2},\cdots,d_{kn'}\}$为点集合中点的度分布;其中要求函数$f_k(d_i)$在作用域$[1,d_{max}]$内单调

递减，其中d_{max}为度分布集合中最高的度数；$f_k(d_i)$可采用线性函数、反比例函数、对数函数等等，本实施例中优选采用$f_k(d_i)$为反比例函数；

[0058] c)每个抽样器均采用遗传算法中的选择算法select(P)，其中P为b)中计算得到的点替换概率集合，选取待替换的点r；选择算法select(P)可采用比例选择算法、确定式采样选择、轮转赌选择等等，本实施例中优选算法select(P)为遗传算法中的比例选择算法；

[0059] d)每个抽样器根据替换原则，判断c)选择的点r是否符合要求，若符合则转至e)，否则转至c)；

[0060] 替换原则为：

[0061] i.上述a)的ii)情况下，在选择被替换点时，不能选择新增边中的点，并且在后续的孤立点删除时也不能删除新增边中的点；

[0062] ii.上述a)的iii)情况下，假定先添加u，再添加v；先添加u时，V_{sk}中没有与其相关联的点，所以发生的替换并没有限制；再添加v时，由于V_{sk}中存在与其相连的点u，所以替换出的点不能为u；在这种情况下，第一个新增点在替换时没有限制，第二新增点在替换时不能替换刚新增的第一个点，并且在后续的孤立点删除时也不能删除刚新增的第一个点。

[0063] e)每个抽样器从V_{sk}中删除r，并从E_{sk}和E_{dk}中删除与r相关联的边；再从V_{sk}中删除独立点；孤立点删除要求不能删除新增点中的第一个点；

[0064] f)每个抽样器把新增点和边加入子图G_{sk}中；

[0065] 其中增加点和边到子图的原则是：如果u,v均在点集V_{sk}中，那么将e=(u,v)加入边集E_{sk}中；如果u,v中一个在点集V_{sk}一个在全局点集合$V_{sglobal}$中，那么将e加入推导边集E_{dk}中；如果u,v均在全局点集合$V_{sglobal}$而不在点集V_{sk}中，不增加点也不增加边到子图中。

[0066] 下面以图2为例，介绍单条边处理流程。单条边处理流程以第i($1\leqslant i\leqslant p$)个抽样器为例，假定当前处理的边为$e_{current}(u,v)$，那么，抽样器i的处理的过程有以下几个步骤：

[0067] 步骤一：如果当前的抽样点数目未达到抽样目标，即$|V_{si}|<n_i$，那么直接选中边$e_{current}(u,v)$，将该边的两个端点u,v分别加入到抽样点集V_{si}中，并将$e_{current}$加入局部抽样边集E_{si}中；

[0068] 步骤二：如果当前抽样点数目已达到抽样目标，那么就要对$e_{current}$进行选择决策。先对$e_{current}$进行基于局部点集V_{si}的边推导，如果u,v都在V_{si}中，那么直接选中$e_{current}$，并将其加入到集合E_{si}中。否则，就要根据全局点集$V_{sglobal}$再对$e_{current}$进行一次推导；

[0069] 步骤三：对$e_{current}$进行基于全局点集推导，如果u,v都在全局点集合$V_{sglobal}$中，那么基于全局点的边推导成功，直接选中$e_{current}$，并将其加入到另一个边集，即全局推导边集E_{di}中。否则，就要对$e_{current}$进行一定概率的选择；

[0070] 步骤四：概率选择后，如果$e_{current}$未被选中，那么直接丢弃$e_{current}$。如果$e_{current}$被选中，那么将$e_{current}$加入到集合E_{si}中，两个端点抽样加入点集V_{si}中，此时至少会产生一次暂存点的替换。在暂存点替换策略上，此处采用的是公平的倒数分布替换策略。此策略中，点集V_{si}中的点被替换的概率与该点的度倒数成正比，这样一来，一定程度上保持了度高的点不会被频繁地替换，又防止了极端地删除底最低的点造成的过度集聚。先利用该策略选出要被替换的点，从V_{si}中删除该点，从E_{si}中删除与此顶点相关的所有边。

[0071] 本领域的技术人员容易理解，以上所述仅为本发明的较佳实施例而已，并不用以限制本发明，凡在本发明的精神和原则之内所作的任何修改、等同替换和改进等，均应包含

在本发明的保护范围之内。

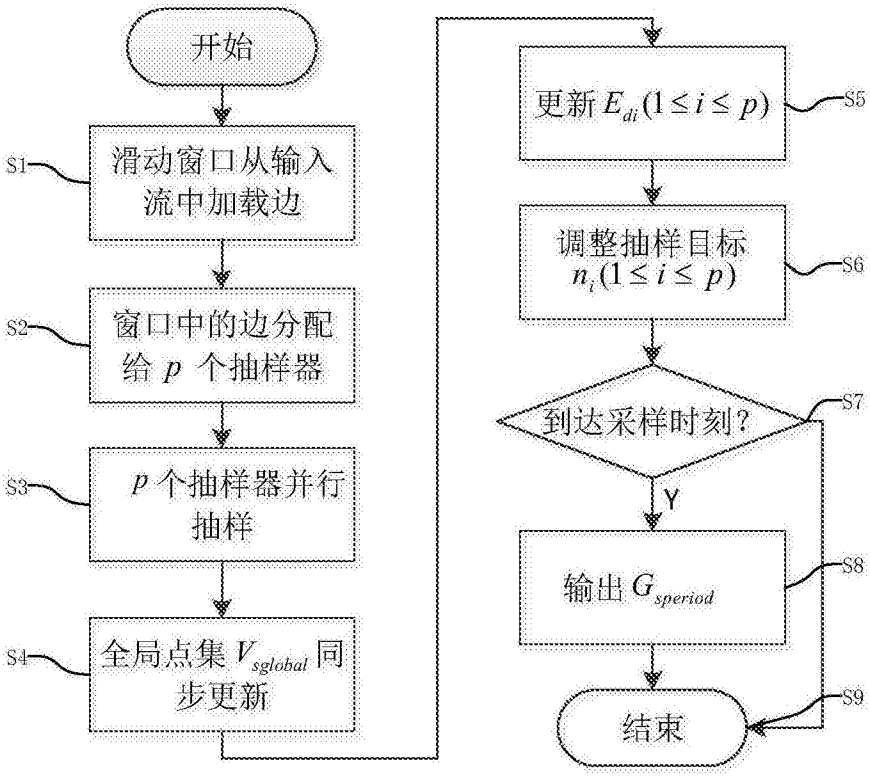

开始
滑动窗口从输入流中加载边
S1
窗口中的边分配给 p 个抽样器
S2
p 个抽样器并行抽样
S3
全局点集 $V_{sglobal}$ 同步更新
S4
更新 $E_{di}(1 \le i \le p)$
S5
调整抽样目标 $n_i(1 \le i \le p)$
S6
到达采样时刻?
S7
Y
输出 $G_{speriod}$
S8
结束
S9

图1

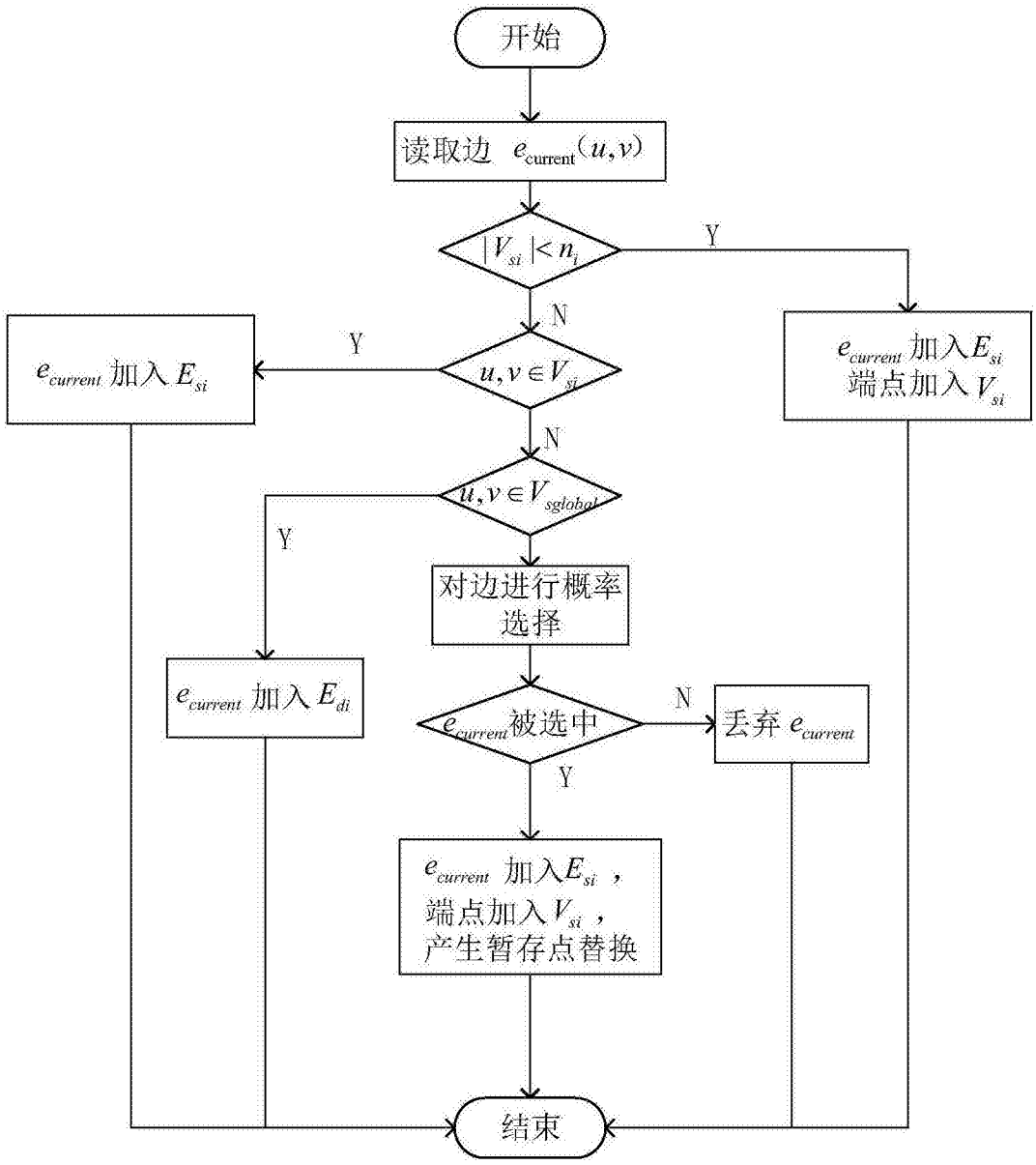

开始
读取边 $e_{current}(u,v)$
$|V_{si}|<n_i$
Y
N
$u,v\in V_{si}$
Y
$e_{current}$ 加入 E_{si}
$e_{current}$ 加入 E_{si}
端点加入 V_{si}
N
$u,v\in V_{sglobal}$
Y
对边进行概率选择
$e_{current}$ 加入 E_{di}
$e_{current}$ 被选中
N
丢弃 $e_{current}$
Y
$e_{current}$ 加入 E_{si} ，
端点加入 V_{si} ，
产生暂存点替换
结束

图2