Network Motif Discovery: A GPU Approach

Abstract

Challenge

> 1. 需要枚举真实图的子图
> 2. 计算大量随机图的所有子图
> 3. 对于数据量几千的生物网络需要几天的时间解决（耗时长）

Motivation

> 利用GPU的高并行度计算**随机图的匹配问题**，进而降低总体时间

Result

> 在生物图中较最好的CPU算法提升两个数量级，以及花费-性能比提升约20倍

I. INTRODUCTION

Motif算法简介

> 对给定图G，以及若干随机图（度数分布同G相似），找到子图g，其在G中出现的频率远大于随机图

现有Motif框架

> Subgraph Technique：对于给定图G，以及k（整数），计数所有顶点个数为k的子图个数
> Frequency Estimation：对于每个子图，计算随机图中该子图出现的期望频率，如果该子图德频率明显高于随机图中平均则返回该子图作为一个motif。

现有问题

Frequency estimation阶段耗时很大，通常需要生成大量随机图（如1000个）计数求平均，这需要大量的子图同构测试，通常现有计数通过加快子图同构测试的速度

该系统相关

> 采用GPU提高并行性，同时考虑了影响GPU性能的三个方面:
> 1.负载平衡
> 2.GPU分支分歧相关操作
> 3.内存访问模式

本篇文章结构

> 1.首先对分析了CPU算法的不足，查明CPU算法移植不到GPU算法的原因,之后采用了一种新的方法有效利用GPU的高度并行性，减轻GPU内核在计算能力方面的限制（Section III/IV/V）
> 2.使用三种优化技术提高可拓展性，避免GPU利用不足，消除冗余计算，使计算开销降低了75%，并且使得能处理之前10倍大的图（Section VI）
> 3.跟之前CPU算法进行比较...（Section VII）

II. PRELIMINARIES

Problem Definition

图结构定义

> 对于有向无标记图$G=(V,E)$，其顶点集记为V，边集记为E，对于任意两点u,v属于V,若存在一条有向边$(u,v)\in E$,u被称为v的$in-neighbor$(相对的v称为u的$out-neighbor$),并且由此定义了出度/入度/bi度,在此基础上对于双向联通的,记为$bi-neighbor$

子图计数定义

> 定义子图$g=(V_g,E_g)$当且仅当至少存在一种映射关系使得$\zeta:V_g\rightarrow V$
> 1.对于任意点$v\in V_g$ 存在$\zeta(v)\in V$
> 2.对于任意边$(u,v)\in E_g$存在$(\zeta(u),\zeta(v))\in E$
> 满足以上要求时判定子图g在G中出现过一次.
> 定义函数$f(g,G)$表示g在G中出现的次数

度分布相似图定义

> 对于图$G^{'}=(V^{'},E^{'})$当且仅当:
> 1.$|V^{'}|=|V|$ 并且$|E^{'}|=|E|$
> 2.存在双射$\psi:V\rightarrow V^{'}$，对于任意$v\in V$,v以及$\psi(v)$有相同出度,入度以及bi-degree

随机图期望定义

> 对于所有度分布相似图组成的集合$\mathcal{G}$,子图g在随机图中出现的概率为:
> $\overline{f}(g)=\frac{1}{|\mathcal{G}|}\sum\limits_{G^{'}\in\mathcal{G}}f(g,G^{'})$
> 由于枚举所有度分布相似图通常不太现实,故实际计算中常用含有r个度分布相似图来代替,上述公式改为:
> $\overline{f}(g)=\frac{1}{r}\sum\limits_{G^{'}\in\mathcal{G}}f(g,G^{'})$

motif定义

> 标准差记为:
> $\widetilde{\sigma}(g)=\sqrt{\frac{1}{r-1}\sum\limits_{G^{'}\in\mathcal{G}_r}(f(g,G^{'})-\tilde{f}(g))^2}$
> 对于用户定义的$\theta(\theta>0)$,对于给定图$G,\mathcal{G}_r,\theta$,当且仅当:
> 1.$\widetilde{\sigma}(g)>0$
> 2.$f(g,G)-\tilde{f}(g)\geq\theta\cdot\widetilde{\sigma}(g)$

问题定义

> 给定$G,r,k,\theta(r>0,k>2,\theta>0)$,找到所有大小为k的motif

符号表

> g : 需要统计的子图
> G :原图
> $\mathcal{G}$:顶点个数为k的所有度分布相似图
> $f(g,G^{'})$:子图g在G^{'}中出现的次数
> $\overline{f}(g)$:子图g的平均出现实际期望
> $\tilde{f}(g)$: 子图g近似的出现期望
> $\widetilde{\sigma}(g)$: 子图g的近似标准差