**北京航空航天大学计算机学院**

**硕士学位论文中期检查报告**

**论文题目**：基于消息量预测的自适应图计算系统性能优化研究

**专 业**：计算机技术

**研究方向**：分布式图计算系统

**研 究 生：**李东泽

**学 号：**ZY1706212

**指导教师：**樊文飞 教授

**北京航空航天大学计算机学院**

2019年 月

目录

[1 论文的研究重点与研究计划 3](#_Toc17155021)

[1.1研究背景与意义 3](#_Toc17155022)

[1.2论文的研究目标 4](#_Toc17155023)

[1.3论文主要研究内容 4](#_Toc17155024)

[1.4 关键技术与难点 4](#_Toc17155025)

[1.5论文预期成果形式 5](#_Toc17155026)

[2 已完成的工作与阶段性成果 5](#_Toc17155027)

[2.1 基于消息量预测的自适应并行模型 5](#_Toc17155028)

[2.1.1 编程模型 5](#_Toc17155029)

[2.2 算法的分类 6](#_Toc17155030)

[2.3 运行时间预测 6](#_Toc17155031)

[2.4 系统的设计与实现 6](#_Toc17155032)

[3 下一阶段工作计划 6](#_Toc17155033)

[3.1 论文研究进度 6](#_Toc17155034)

[3.2 尚未完成工作 7](#_Toc17155035)

[3.3 下一阶段计划 7](#_Toc17155036)

[4 主要参考文献 7](#_Toc17155037)

基于消息量预测的自适应图计算系统性能优化研究

# 1 论文的研究重点与研究计划

## 1.1研究背景与意义

图计算是以“图论”为基础的对现实世界“图”结构的抽象表达。在实际的应用中，存在着许多图计算问题，如图的连通性[1]、单源最短路径[2]、网页排序[3]等，同样在社会或科学领域中的许多问题都可以通过转换成图计算模型从而得到解决，例如: Google公司每天都需要对数亿级别的网页进行排序[4]; 生物医学上需要对蛋白质进行子图匹配，分析蛋白质间的相互作用从而开发出更有效的临床药物[5]。因此，对图数据的分析和计算具有很重要的现实意义。

早期的图计算数据规模较小，大部分使用BGL等单机图处理算法[6]，极大的限制了图计算问题的规模。后来随着图规模的增长，出现了Parallel BGL[7]等图计算系统，但是系统本身不具备良好的稳定性。近些年来，随着大数据时代的来临，大图数据的分析给图计算带来了极大的挑战。于此同时，MapReduce[8]作为一个并行的计算框架，及其开源实现Hadoop[9]为海量数据的处理提供了便利，但是由于图计算过程中通常涉及多轮迭代，因此该框架不能高效的进行并行图计算。

针对上述图计算特点与图处理中所面临的问题，面向大规模图计算的研究吸引了越来越多的研究者投身其中。最近，在计算模型和运行优化方面涌现出不少具有前瞻性的图计算系统，代表性的图系统包括Giraph[10]、GPS[11]、GRACE[12]、GraphLab[13]、PowerSwitch[14]、GRAPE+[15]。它们在计算模型上分别属于同步、异步、混合以及自适应计算模型，它们围绕着减少迭代次数加快算法收敛、减少消息传递数目、实现负载均衡这3个优化目标，分别采用一种或多种优化技术对自身进行优化，但是它们往往都是在优化某个目标的同时牺牲了其它性能。

因此本课题以自适应并行图计算为框架，采用机器学习模型，针对图计算过程中接收的消息量进行预测，最终合理的动态调节各计算节点间的相对进度，从而达到性能优化。

## 1.2论文的研究目标

在已有的图计算模型中，同步模型存在“木桶效应”问题；异步模型存在大量的冗余计算；混合计算模型为实现同步与异步间自由切换产生了额外的内存与预测开销；而自适应模型中，每个计算节点通过引入限定值动态调整节点间相对进度，从而解决了同步、异步、混合模型下的问题，本文在已有自适应计算模型的原型下，提出了新的基于消息量预测的自适应图计算模型，该模型可在慢机的情况下规避上述模型的弊端，加速并行图计算。

## 1.3论文主要研究内容

本文研究重点与研究内容如下：

   1）提出基于消息量预测的自适应并行图计算模型，以下简称 MPAP模型（Message Predict Async Parallel），并将对消息量的预测转化为对运行时间及消息到达速率的预测。

   2）我们将对算法运行时间的预测问题定义为机器学习中的回归问题，使用均方相对误差（MSRE）做为损失函数，最终选取机器学习中适合图计算领域的回归模型作为候选，进行训练预测，从而分析各模型在分布式图计算时间预测问题上的优缺点。

   3）我们选取包含单源最短路径（SSSP）、深度优先搜索（BFS）、网页排序（PageRank）、节点结构分析（HITS）、图采样（Graph Sampling）、图连通性（WCC WCC-HASHMIN）、标签传播（Label Propagation）、社区发现算法（Louvain）等9种学术届与工业界常用的图计算应用算法，并尝试将算法进行分类，针对每类算法，给出不同的特征提取方案，并展示预测效果。

   4）我们在已有的自适应图计算系统原型下[GRAPE+]，开发完成基于消息量预测的自适应并行图计算系统，并与同步、异步、及自适应原型的系统进行对比，评估效果。

## 1.4 关键技术与难点

(1) **图应用算法的分类与特征选取**：收集运行时信息的过程及机器学习领域特征提取的过程，特征提取结果的好坏直接影响接下来模型的训练与预测，图计算中的特征提取涉及多方面因素，包括图结构本身，如图的规模、平均度数、各子图边界点个数；运行时信息，如消息的接收数量与质量等不同特征。本课题从分布式图计算角度出发，会对上述算法从消息传递、本地计算以及图数据特性等不同方向分析，从而将图算法进行分类，给出每一类算法的特征选取方向与不同模型的预测效果对比。

(3) **模型的选取**：虽然我们可以将运行时间的预测定义成机器学习中的回归问题，但存在众多的回归模型需要我们进行分析，如局部加权线性回归模型、岭回归模型、树回归模型甚至使用神经网络模型，不同模型在训练的时间及效果上差异很大，因此结合图计算领域约束，我们期望找出每类算法适合的回归模型，并作出训练与预测。

(4) **图计算系统实现**：最终，我们需要将训练好的模型嵌入分布式图计算系统中，嵌入过程主要包括运行时日志采集、特征提取、模型训练及预测，因此整个过程涉及分布式图计算系统的图存储模块、消息传递模块、及与模型的训练预测有关的自适应调整模块的开发。

## 1.5论文预期成果形式

(1) 给出基于消息量预测的MPAP模型。

(2) 针对多种工业界与学术界常用的图应用算法，在PIE[]模型下，找到不同算法的运行时特点，并对算法进行分类，针对每一类算法给出对应特征提取方案与不用模型的训练效果对比。

(3) 基于已有的自适应图计算系统，开发基于消息量预测的自适应并行图计算系统，并与同步、异步、自适应原型模型进行实验对比，期望运行时间较同步、异步、自适应原型系统相比，有一定的缩短，消息量减少。

(4) 基于后段图计算服务，开发web页面，对整个消息量预测的图计算流程进行功能性展示。

# 2 已完成的工作与阶段性成果

## 2.1 基于消息量预测的自适应并行模型

### 2.1.1 PIE编程模型

由于MPAP采用与[GRAPE] 相同的编程模型，即PIE模型，因此本节首先简要介绍一下该模型。

**图符号术语**：给定某有向图或无向图，其中代表图中点的有限集合，代表图中边的集合，代表图中顶点或边上的属性；

给定某一自然数，我们采用某种分图策略将图分割为个片段，其中每个均为的子图，且有，，，，针对每个顶点 ，有，针对每条边，有。

MPAP模型采用边分割或点分割策略的来分割图，当使用边分割时:

(1) 代表点的集合，其中针对每个点存在一条边 ，使得，且。

(2) 代表点的集合，其中针对每个点存在一条边 ，使得，且。

我们将上述成为子图的边界点集合

给出一个sample图，举个例子，之后给出相应的术语表(PEval)

**PIE编程模型**：众所周知，图计算问题作为图查询问题的一类，针对某一问题实例给定问题，为回答该问题，MPAP模型需要用户提供以下三个函数：

(1) PEval (局部估值函数): 该函数将子图与问题作为输入，并在子图上计算回答该问题，得到结果，其中针对该问题的所有单机图算法均可直接使用。

(2) IncEval (迭代增量函数): 该函数将已有结果 、子图增量 、与问题作为输入，并在更新后的子图上回答问题，保证有 ，其中 是在原有的基础上做的更新计算。

(3) Assemble (聚合函数): 该函数用于聚合其他子图在PEval或IncEval阶段计算所产生并同步本子图的更新量，最后将聚合结果更新到已有结果。

## 2.1.2 MPAP 模型

本节我们介绍基于消息量预测的并行图计算模型，即MPAP模型

**初始配置**：MPAP模型本质上仍然采用上节介绍的GRAPE PIE编程模型，给定某一图计算问题实例及在该实例上的问题，MPAP模型需要用户指定上述PIE函数。首先，将输入的图数据 通过边分割或点分割策略分为多个片段，每个片段存储在某台虚拟机上（术语worker），除此之外，MPAP模型需要做为coordinator节点，用于完成后续将要提到的集群终止状态检测与运行时特征的收集、训练和预测任务。因此，若 ，则多个片段会分到同一个虚拟机上并分享内存。

**并行模型**：与PIE 模型不同的是，MPAP模型还需增加以下声明，其中所有在PEval中的声明均在IncEval中共享。

1. 状态变量：PEval中每个片段都会声明并维护一个集合，该集合用于存储片段中每个点的状态信息，其中包括结果信息，在接下来的IncEval阶段中，该集合也用于更新有用的消息。
2. 聚合函数：PEval中需要指定聚合函数来解决多台机器传递的消息同时作用于相同的变量的问题，常用的聚合函数如函数、函数、函数等。
3. 特征提取：IncEval中每个片段在接收处理本轮的消息后，会向coordinator节点发送本轮的运行时日志，该日志主要包含当前该片段的状态信息与特征信息。

下面我们以单源最短路径问题为例，给出对应的MPAP并行模型

单源最短路径问题中，给定一个带权有向图 ，其中代表每条边上的权重且，给定某点做为源点，现要计算从源点到图中其它各顶点的最短路径长度，该长度即为到目标顶点的路径中所走边的权重之和，该算法输入输出表述如下:

Input: 带权有向图 ，与图中某顶点

Output:

|  |
| --- |
|  |

如上表所示，MPAP在PEval阶段中采用了众所周知的求单源最短路径的算法- Dijstra算法。但与该算法单机版本不同的是1) 我们针对该fragment中的每个点定义元组，其中代表该点与原点的距离，初始化为无穷大，意味着与原点不可达；代表在本轮计算中该点是否被访问过，初始化为。该元组声明在PEval中，但也在IncEval中共享，后续也同样用于更新IncEval接受来自其它worker的消息。2) 我们采用做为聚合函数，即当该fragment中某个点接收到了来自多个worker的更新消息，我们选取其中的最小值做为最终的唯一更新消息。

上述PEval过程执行结束后，每个fragment会将外部点的距离发送到对应的worker上，MPAP消息传递的过程下一节详细阐述。

|  |
| --- |
| *（加）* |

如上表所示IncEval过程首先对接收的消息执行聚合函数，并更新对应点的值，之后同样采用Dijstra算法计算本轮最短路径，并将计算后的最新值以消息的方式传递给其它worker。与原有AAP(是不是要介绍一下)模型不同的是，IncEval过程还需将本轮计算的运行时信息发送给coordinator所在的节点，用于接下里的模型训练与预测，详细内容下节阐述。

|  |
| --- |
| *min step*  *if* |

除PEval与IncEval函数外，MPAP模型还需要Attl函数，用于动态调节各worker间相对进程，该函数对用户透明，具体细节后续给出。此例中不需要Assemble函数，因此不做介绍。

**消息传递**：

**动态开关**：

**终止条件**：

**整体流程**：

## 2.2 算法的分类

## 2.3 运行时间预测

## 2.4 系统的设计与实现

# 3 下一阶段工作计划

## 3.1 论文研究进度

## 3.2 尚未完成工作

## 3.3 下一阶段计划

# 4 主要参考文献