**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。

**（状态示意图）**

**消息传递**：每轮计算结束后，每个worker都会收集属于自己的fragment的更新结果，这些结果来自与每个fragment内部更新数据的改变量，并将更新结果发送给其他的worker，为实现该消息传递机制，每个worker需声明维护以下数据结构：

1. 图中全部点与该点所属worker的映射，以便顺利将该点的更新消息放松给对应worker
2. 一个缓冲buffer，用于接收来自其他worker发送的消息

MPAP的消息传递过与经典AAP模型相同，是点对点通信、且发生在图计算的任意时刻，即任意worker pi在任意适合都可将消息发送给worker pj而不需要关系pj此时处于状态，同样，worker pi也会在任意时刻接受来自其他worker的消息，保存在缓冲buffer中，该过程不会阻塞任何计算过程。但与AAP模型不同的是，每个fragment除发送本轮更新的消息外，还需像coordicator节点发送本轮的运行时信息，该信息用于后续的模型训练。

**MPAP模型**：MPAP模型中，为解决同步模型下的“木桶效应”及异步模型下的冗余计算问题，每个计算节点引入了限定值，之后每轮迭代计算开始前，各计算节点需判断是否等待长的时间以积累更多的消息，我们给出基于下述函数进行动态调整：

上述函数中的各参数描述如下，表示某一计算节点：

(1) ：表示计算节点当前接收的消息量。直观上来看，如果 越大，则当前计算节点接收消息越多，应尽快开始下轮迭代计算。

(2) ：表示本轮迭代计算开始前期望接收的消息量。该变量定义如下：，其中可由用户自定义初始化，为了尽快减少冗余计算；为消息到达速率；是 的一部分，后者为运行时间。

(3) ：表示计算节点为接收到期望的消息量而等待的时间，定义为；表示计算节点前一轮计算后的空闲时间，以防止无限期的等待。

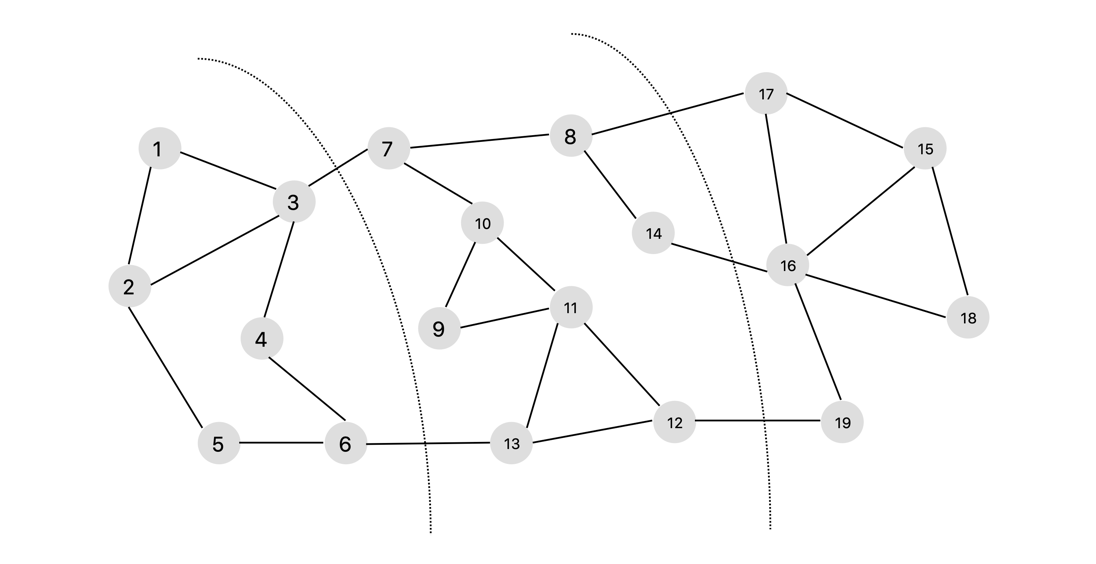
**终止条件**：MPAP模型的终止条件与经典AAP模型相同，即新一轮迭代计算开始前，如果当前没有任何接收到的消息，则会像coordinator发送inactive flag, 当coordinator接收到所有worker的inactive flag后，会广播各个worker 终止命令，各个worker会对此返回ack确认自己是否inactive，如果某个worker仍有任务需要计算，则回应wait命令，继续进行下一次计算，coordicator重新接入监听等待状态。

**整体流程**：

## 2.2 算法的分类

本章讨论了学术届与工业届常用的9种经典图应用算法，包括 sssp、bfs、pagerank、wcc、sampling、hits、label propagation、wcc\_hashmin、louvain, 我们发现在算法运行过程中，通常一次迭代不需要访问所有的内部点，因此在PIE模型下，我们调查了上述算法的运行时行为，并将行为定义为每一次超步实际访问的活跃点的集合，之后我们调查了不同算法在一次图任务计算中该集合的变化情况，最后我们将上述图算法分为三类，针对每类我们给出了不同的图特征提取方案，并采用不同学习模型进行训练预测。实验表明，我们我们预测的效果十分准确。

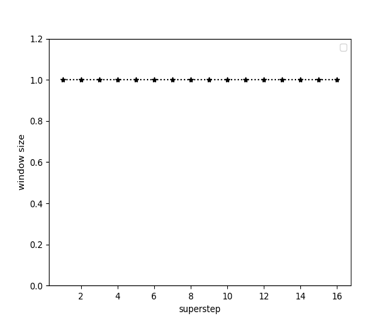
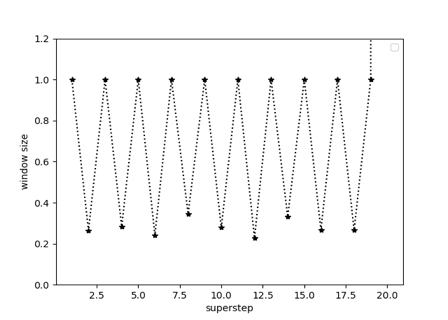
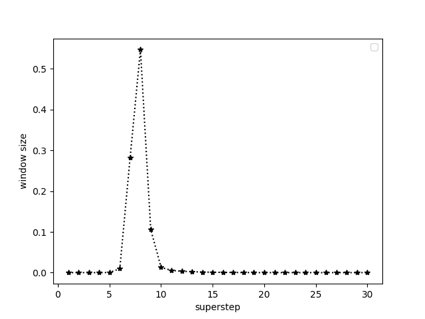
1. Active Nodes Window 行为

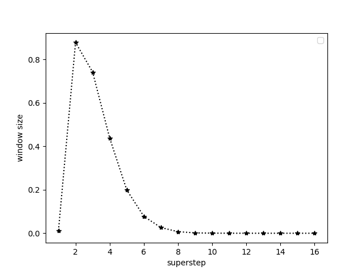
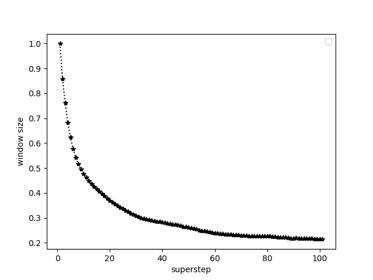
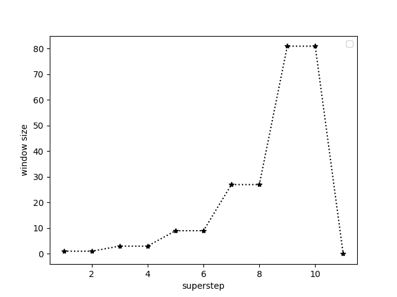


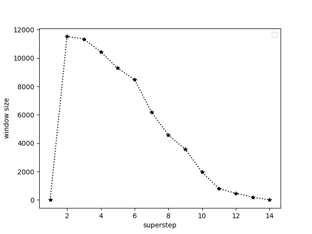
我们以小的图数据为例，介绍不同种类图算法的Active Node window的变化情况。如图所示，图包含19个节点以及26条边，假定该图采用边分割的策略分为3个子图F1，F2，F3,其中F1包含点v1={v1,v2,v3,v4,v5,v6}以及与这些点所连接的边，F2包含点V2={v7,v8,v9,10,v11,v12,v13,v14}以及与这些点所连接的边，F3包含点V3={v15,v16,v17,v18,v19}以及与这些点所连接的边其中F1与F2有两条边相连，F2与F3有3条边相连。

图2展示了pagerank，sssp以及hits算法在5个超步运行下的active node window情况，每个超步中，浅灰色的点代表非活跃点，深灰色点代表活跃点；其中图2-1展示了pagrank在前5轮超步下的working window，图中所有的点在每轮计算中均为活跃点，同样每个Fragment均参与了每轮计算；图2-2展示了sssp在前5轮超步中下的working window，可以看出5轮迭代计算的working window分别是W1={},W2={},W3={},W4={},W5={}, 并且前两轮计算，只有F1处于活跃状态，第三轮是F1，F2处于活跃状态，但是整体的active node数目很少，从4轮迭代之后，F1,F2,F3均处于活跃状态，且整体active node数目增多；图2-3代表了hits算法在前5轮超步下的working window,可以看出该算法在第一，第三，第5轮迭代中，所有点均为活跃点，而第二，第四轮迭代中只有部分点（接收更新小消息的点）处于活跃状态。

因此我们可以看出，不同的图算法在每轮迭代计算中，active nodes有不同的变化效果，接下来我们给出所有算法的现实中图数据上的active node变化折线图，所有算法都需要实现上述用户定义的PIE函数。







从图中可以看出，pagerank，label-propagation算法在每轮计算中，所有点均处于active状态，因此我们将该类算法称为Always Active Style, 而sssp bfs sampling 算法的active node window在折线图上表现为先升高，后降低（其中sampling之后采样终止时降为最低），wcc wcc-hashmin算法active node window在折线图的表现为开始升高，之后逐渐下降，我们将上述5种算法均成为Message Dependency Style ，因为每轮迭代计算活跃点的总数随着以来与当前Fragment接收到的消息数，（后面我们会给出不同种类的算法如何指导我们抽取不同的图计算运行时特征），hits 以及louvain算法active node window 表现出一定的规律性。其中hits算法在奇数轮迭代计算中，所有的点均处于active 状态，而偶数轮的迭代计算中部分点处于active状态，这类算法典型的操作一次计算流程分为多个超步，因为我们将这些算法成为Multiple Parse Style，这类在图算中也十分常见，包括MST MM等。

**Always Active Style (种类一)**

在该风格下，每个子图在每一次超步计算中，都需要遍历所有的内部点，完成规定的图计算任务，并将本轮计算生成的消息传给邻居fragment（前面要讲述一下fragment），表算法-1描述了该类算法在IncEval上的行为，其中由于PIE模型下的Peval只在图计算开始时执行一次，因此本文只考虑IncEval的内容

表

我们可以看出，输入M代表该fragment接收到来自于上一轮迭代其它fragment发来的消息，在IncEval中，首先针对接收到的消息调用聚合函数，并将消息更新到本fragment的内部点中（多少行），其次针对本fragment的所有内部点进行相应计算（多少行）最后将更新后的消息发往其他的fragment（多少行）。

作为一个例子我们给出PIE模型下pagerank的IncEval计算过程，从算法中我们清楚的看到，这看看能不能用论文的那里

**Message Dependency Style (种类二)**

在该风格下，通常计算开始时只有一个点或若干点处于活跃状态，而其它点是否参与本次计算取决于是否接到到消息，因此消息的传播条件以及决定了这类算法的活跃点数目，如果算法-4所示，开始条件下，只有一个点处于活跃状态，其余点均为非活跃点，在IncEval开始时，首先针对接收到的消息调用聚合函数，并将消息更新到本fragment的内部点中（多少行），同时记录更新的内部点，标记为active状态，接下里，针对本fragment活跃的内部点，进行相应计算，而没有更新消息的内部点仍处于非活跃状态，不参与本轮迭代计算，最终将更新后的消息发往其他的fragment。跟always对比一下。

给出sssp的例子

**Multiple Phase Style (种类三)**

如算法-5中所描述，该类图算法无法在一次超步中完成一次图任务计算，例如hits算法，该算法与pagerank算法类似，都是对一个网络图做结构分析，最早用于搜索，现在也常用来做社交网络结构分析，但pagerank算法不同的是，hits算法一个(网络)节点的重要性分成了两种属性，分别为Authority于Hub，其中Authority认为如果一个页面提供了关于某个主题的信息，那么这个页面就是有价值的，这样的页面就是authority页面，因此如果有很多其它的页面都指向该页面，说明这个页面的authority值就高；Hub任务还有一些页面，比如google页面，虽然它本身不提供任何主题信息，但是从这个页面出发，可以跳转到很多有价值的页面上去，这样的页面就会成为hubs，因此如果某个页面指向了很多有authority值高的页面，那么它的hub值就会高。在PIE计算模型中，因为一轮中，每个点既要算本身的authority值，也要算hub值，并且hub值的计算要依赖于周围邻居节点的aythority值，所以无法用PIE模型一轮的IncEval同时算出所有点的authority值与hub值，因此需要拆分成两次IncEval去做，其中一次用于计算所有的Authority值，另一次根据上轮的authority值计算hub值。通常情况下，这类算法会将一次图计算任务Pi分为PIE模型下的n次IncEval：Pi0…Pin-1，这类算法还包括louvain，mst。

我们以hits算法为例。。。

## 2.3 运行时间预测

在PIE计算模型中，某个fragment（术语表）每轮的计算开销主要依赖于活跃点的数目与接收到的消息数。其中活跃的点是该fragment内部点的子集，因为在某些算法中，并不是所有的点都会被接收到的消息激活，我们将某个fragment Fi在第s次IncEval迭代计算的活跃点集合记为Wis，因此在第s次计算中，Fi的计算开销可以表示为Ks=wis + outnode +msgnum，（解释一下）

等式

上述等式也同样表达了，针对一轮计算，某个fragment的执行时间取决以下三个要素

## 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