**中图分类号：**

**论文编号：**



硕 士 学 位 论 文

**基于运行时间预测的并行图计算系统性能优化研究**

作者姓名 李东泽

学科专业 计算机技术

指导教师 樊文飞 教授

培养院系 计算机学院

**The Research and Implementation of Email Importance Evaluation Method based on Thread Network**

A Dissertation Submitted for the Degree of Master

**Candidate：Li Dongze**

**Supervisor：Associate Prof. Fan Wenfei**

School of Computer Science and Engineering

Beihang University, Beijing, China

**中图分类号：**

**论文编号：**

硕 士 学 位 论 文

基于运行时间预测的并行图计算系统性能优化研究

作者姓名 李东泽 申请学位级别 工学硕士

指导教师姓名 樊文飞 职 称： 教授

学科专业 计算技术 研究方向： 分布式图计算

学习时间自 年 月 日 起至 年 月 日 止

论文提交日期 年 月 日 论文答辩日期 年 月 日

学位授予单位 北京航空航天大学 学位授予日期 年 月 日

关于学位论文的独创性声明

本人郑重声明：所呈交的论文是本人在指导教师指导下独立进行研究工作所取得的成果，论文中有关资料和数据是实事求是的。尽我所知，除文中已经加以标注和致谢外，本论文不包含其他人已经发表或撰写的研究成果，也不包含本人或他人为获得北京航空航天大学或其它教育机构的学位或学历证书而使用过的材料。与我一同工作的同志对研究所做的任何贡献均已在论文中作出了明确的说明。

若有不实之处，本人愿意承担相关法律责任。

学位论文作者签名： 日期： 年 月 日

学位论文使用授权书

本人完全同意北京航空航天大学有权使用本学位论文（包括但不限于其印刷版和电子版），使用方式包括但不限于：保留学位论文，按规定向国家有关部门（机构）送交学位论文，以学术交流为目的赠送和交换学位论文，允许学位论文被查阅、借阅和复印，将学位论文的全部或部分内容编入有关数据库进行检索，采用影印、缩印或其他复制手段保存学位论文。

保密学位论文在解密后的使用授权同上。

学位论文作者签名： 日期： 年 月 日

指导教师签名： 日期： 年 月 日

摘 要

在社交网络服务走向多元化的今天，邮件由于其良好的隐私性和普及性仍然拥有庞大的用户群体。

**关键词：**邮件分类，特征评价，邮件会话网络，EmailRank

**Abstract**

Nowadays, with the diversification of social networking services, the email still maintains a large user base due to its good privacy and popularity.

**Key words:** email classfication, feature evaluation, email thread network, EmailRank

目 录

[摘 要 I](#_Toc23948477)

[**Abstract** II](#_Toc23948478)

[目 录 III](#_Toc23948479)

[图 目 V](#_Toc23948480)

[表 目 VI](#_Toc23948481)

[第一章 绪论 1](#_Toc23948482)

[1.1 研究背景及意义 1](#_Toc23948483)

[1.2 国内外研究现状 1](#_Toc23948484)

[1.2.1 邮件重要性定义 1](#_Toc23948485)

[1.2.2 研究现状小结 1](#_Toc23948486)

[1.3 研究目标和内容 1](#_Toc23948487)

[1.4 论文组织结构 1](#_Toc23948488)

[第二章 相关理论与技术研究 2](#_Toc23948489)

[2.1 邮件重要性评价方法 2](#_Toc23948490)

[2.1.1 基于社交网络的邮件重要性评价 2](#_Toc23948491)

[2.1.2 基于邮件特征的邮件重要性评价 2](#_Toc23948492)

[2.2 邮件会话网络的构建 2](#_Toc23948493)

[2.2.1 基于元数据的会话网络构建 2](#_Toc23948494)

[2.3 社交网络中节点重要性评价 2](#_Toc23948495)

[2.3.1 基于网络拓扑的节点重要性评价 2](#_Toc23948496)

[2.3.2 基于交互信息的节点重要性评价 2](#_Toc23948497)

[2.4 其他技术 2](#_Toc23948498)

[2.5 本章小结 2](#_Toc23948499)

[第三章 邮件重要性特征分析 3](#_Toc23948500)

[3.1 邮件特征 3](#_Toc23948501)

[3.2 特征有效性实验 3](#_Toc23948502)

[3.2.1 数据集及标注 3](#_Toc23948503)

[3.2.2 分类器选择 3](#_Toc23948504)

[第四章 邮件会话网络的构建 4](#_Toc23948505)

[4.1 邮件会话网络的改进 4](#_Toc23948506)

[4.2 邮件会话网络的生成 4](#_Toc23948507)

[4.3 邮件会话网络的存储 4](#_Toc23948508)

[4.4 实验评估 4](#_Toc23948509)

[4.4.1 实验方案 4](#_Toc23948510)

[第五章 基于会话网络的邮件重要性评价方法 5](#_Toc23948511)

[5.1 PageRank算法 5](#_Toc23948512)

[5.1.1 实验方案 5](#_Toc23948513)

[第六章 EmailRank算法在实际系统中的应用 6](#_Toc23948514)

[6.1 应用于邮件客户端 6](#_Toc23948515)

[6.1.1 需求分析 6](#_Toc23948516)

[6.2 本章小结 6](#_Toc23948517)

[总结与展望 7](#_Toc23948518)

[论文总结 7](#_Toc23948519)

[未来工作展望 7](#_Toc23948520)

[参考文献 8](#_Toc23948521)

[攻读硕士学位期间取得的学术成果 9](#_Toc23948522)

[致 谢 10](#_Toc23948523)

图 目

[图 1邮件信息过载问题是一项价值6500亿美金的经济累赘 1](#_Toc531617249)

[图 2 ERIC邮件客户端的用户界面 4](#_Toc531617250)

[图 3信任网络结构 8](#_Toc531617251)

[图 4 EMIRT算法信任分数计算原理图 9](#_Toc531617252)

[图 5重要邮件分类方法的架构 11](#_Toc531617253)

[图 6邮件会话的两种网络结构 13](#_Toc531617254)

[图 7 Webmagic系统架构 18](#_Toc531617255)

[图 8 Lucene的架构和全文检索过程 19](#_Toc531617256)

[图 9 Spring Data Neo4j的应用架构 21](#_Toc531617257)

[图 10不同分类器实验效果对比 27](#_Toc531617258)

[图 11基于特征方法与基准方法分类效果对比 28](#_Toc531617259)

[图 12邮件的两种引用方式 31](#_Toc531617260)

[图 13改进后邮件会话网络构建示意图 32](#_Toc531617261)

[图 14邮件关系匹配流程图 34](#_Toc531617262)

[图 15 Neo4j浏览器界面中的邮件会话网络 36](#_Toc531617263)

[图 16改进和未改进会话网络对比图 36](file:///C:\Users\jiang\Desktop\最终论文\毕业论文_蒋坤_SY1606301.docx#_Toc531617264)

[图 17某用户邮件关系网络中的重要节点识别 44](#_Toc531617265)

[图 18 EmailRank和基于邮件特征方法实验结果对比 45](#_Toc531617266)

[图 19原始Nylas-Mail客户端界面 48](#_Toc531617267)

[图 20插件ThreadNetwork效果图 48](#_Toc531617268)

[图 21 W3C会员影响力评价系统架构图 50](#_Toc531617269)

[图 22分布式爬虫架构图 51](#_Toc531617270)

[图 23 W3C会员影响力评价系统截图：影响力排名 53](#_Toc531617271)

[图 24 W3C会员影响力评价系统截图：影响力对比 53](#_Toc531617272)

表 目

[表 1节点中心度的度量方法 16](#_Toc531563451)

[表 2四类邮件特征列表 22](#_Toc531563452)

[表 3 SemanticDistance的算法描述 24](#_Toc531563453)

[表 4邮件重要性特征重要权值 30](#_Toc531563454)

[表 5改进和未改进会话网络下的方法性能对比 37](#_Toc531563455)

[表 6 EmailRank和基于网络拓扑方法的实验结果对比 45](#_Toc531563456)

# 绪论

## 研究背景及意义

在社交媒体、社交网络和即时通讯飞速发展的如今，随着“互联网+”以及各式互联网应用等新型产品，如抖音、猿辅导的兴起和云计算等新相关技术的飞速发展，人类获取数据的途径和规模正以前所未有的速度快速增长，于此同时，与大数据相关的技术变革于理论发展成为了当今世界的热点话题。大数据在环境科学、超级计算、生物学、互联网经济学等领域以及通讯、航天、金融等行业普遍存在，并引起人们的广泛关注。在这些领域或行业中所涉及的数据通常具有某种关联性，从而产生了大图数据。图作为计算机数据结构中一种基本且常用的数据类型，是表示实体与实体间关系的方法，一张图由若干顶点和顶点间所连接的边组成，其中根据连接边是否具有方向的不同又分为有向图与无向图。无论哪种表示方式，其在表达的语义甚至结构方面都比线性表和树更为复杂。社会或科学领域中的许多应用场景都可以通过转化成图计算模型从而得到解决，与图相关的处理和应用几乎无所不在。比如，社会安全部门需要对整个社交网络进行精准的模式匹配，从而增加寻找贩毒、犯罪集群的可能性；交通运输部门需要确定最优的运输路线；生物信息领域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[1]。不仅如此，互联网行业中的用户账号匹配、社交网络分析、风险控制模型的训练都可以看作是大图数据的应用。因此，都大图数据的分析与计算具有很重要的现实意义。

早期传统的图数据管理与图分析计算技术通常只针对“小图”进行处理，整个计算过程中不需要过多甚至复杂的迭代过程，更不会产生大量需要信息通讯的消息，时间与空间的开销相对较低。但是，现实世界中实体规模的持续扩大，导致相应图模型的数据规模迅速增长，甚至有数十亿顶点以及上万亿条边的图数据，而这些只靠顶点和边数量组成的图数据结构还只是大图数据的冰山一角。很多应用中涉及的复杂图数据为了表达其现实语义，往往在顶点以及边上附加一定的属性信息，如表示人的节点通常含有姓名、电话、地址等相关属性，这些属性的存储与计算代价通常是点和边的数十倍乃至数百倍。而且，相比于图数据上的搜索查询，大图上的分析计算算法往往需要基于图的结构进行递归迭代操作，直至满足收敛条件或达到预设的最大迭代轮数，因此会产生巨大的通讯数据，并需要频繁的处理中间结果。此外，除了静态数据与离线计算外，很多情况下实时处理情景需要图数据是动态变化的，如淘宝“双十一”实时分析风控数据等，这种动态往往以流式载图作为扩图途径，不断改变着大图数据的规模和结构，因此面对如此大规模的静态和动态图数据，其底层存储、索引建立、统计分析、实时计算等处理的时间开销和空间开销远远超出了早期传统图处理的能力，因此如何解决大图数据的存储技术与查询计算技术，已经成为数据库领域需要迫切解决的问题，也是大数据领域极具挑战的工作。

面对上述大图数据存储与计算的诸多问题，研究者们已经不满足于仅通过传统的方法探讨基本的图管理与分析问题，一些新兴的研究热点也在不断涌现，如图查询语言、图计算框架等。对于这些研究方向，虽然早期也有一些相关的探讨，但随着计算机软硬件的以及云技术的不断发展，如何提出更准确高效的方法来解决这些问题，为图管理与分析工作带来了新的机遇与挑战。Parallel BGL[7]最为最早一代的并行图计算系统，其在经典的图应用算法基础上开发了并行版本，但此类系统大多数本身不具备一定的稳定性与容错性，并且部分图应用算法的正确性也得不到保证，为了解决稳定性的问题，一部分研究团队专门为特定的图计算应用开发了分布式图计算架构来加速图数据的并行计算，虽然针对性的架构处理可对特定图应用算法进行深度的代码优化以达到更好的处理性能，但毕竟过于局限，针对某一新的图应用算法，算法的移植甚至重开发操作需要重写大量的代码，不仅代码的复用率极低，框架的通用型、可拓展性以及可维护性也很差，对编程人员与使用人员的挑战性极大；MapReduce[8]作为一个并行的计算框架，及其开源实现Hadoop[9]为海量的数据处理提供了便利，为此，用户只需要定义Map和Reduce函数就可完成相应的逻辑计算，不需要关心并发、容错、一致性等复杂问题。但是由于MapReduce框架的自身特点，任务的执行划分为多次Map和Reduce阶段，每个阶段都需要多次的读写分布式文件系统，并且，任务的执行期间无法共享更新数据，再加上图计算的过程通常涉及多轮的迭代，各迭代涉及顶点间的复杂联系，且迭代完成时还要依赖对点的消息传递以满足图计算正确性，因此若将每轮迭代抽象成MapReduce作业不能高效的进行并行图计算，虽然Haloop[10]等处理框架对MapReduce过程进行许多改进，但是并没有从根本上解决基于分布式文件系统所带来的I/O开销问题。针对上述图计算框架的问题，为解决对大规模图数据计算的迫切需要和传统的计算平台无法满足当今的图处理需求的问题，面向大规模并行图计算框架的研究吸引了越来越多的研究者投身其中。

。

## 国内外研究现状

目前面向大规模图数据计算的分布式并行图计算模型及其相关优化研究取得了不错的结果，依据调度方式的不同，大致分为同步计算模型、异步计算模型、混合计算模型、自适应模型四类。

### 同步计算模型

为改进MapReduce计算模型在大规模图数据上的计算效率低等问题，1990年，图领奖获得者Valiant提出了基于消息传递的BSP(bulk synchronous parallel)计算模型[11]，Google也在2010年推出了基于该计算模型的大规模图数据并行计算框架Pregel[11]，这种计算模型区分于现有的MapReduce计算框架在于其适合做数据的迭代计算，其采用一种“块”同步的思想，即通过消息传递机制实现块间数据同步、块内并行计算。其核心思想是将每一次迭代称为一个超步(Superstep)，整个图计算划分为多个超步，而一个BSP作业由一系列顺序执行的并通过全局同步路障分开的超步组成，并行任务时每个节点按照各自的超步组织，接收来自上一个超步的全局消息，执行本地计算并将本次计算生成的消息发送给下一次超步，计算过程如图2.1所示，并行迭代直到全部顶点收敛，即每个顶点相邻两次迭代的变化量小于给定的常量或达到预设的最大迭代次数。

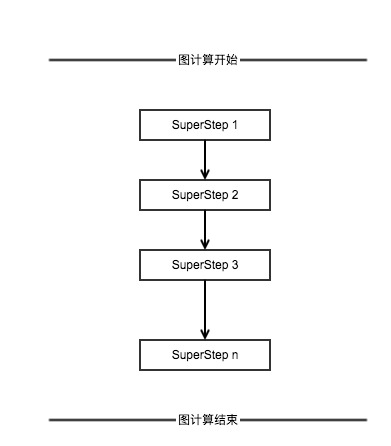


图2.1 图计算垂直结构

在每一次的超步计算中（图体现出超步的内部计算），又会分为本地计算、消息传递、全局同步三个阶段，图2.2、2.3（这两个这图感觉不太好）分别展示了每个超步过程中的计算过程及同步模型的整体计算流程。

1. 本地计算: 计算节点间相互独立，每个节点负责读取本地内存数据执行并行计算。

2. 消息传递: 本地计算阶段完成后，每个计算节点将生成消息通过通信传递给与之相关联的其他节点，以便交换彼此计算所需要的数据。

3. 全局同步: 用于数据间的整体同步，即块见同步。各个计算节点在完成本地任务后进入全局同步阶段，先完成本地任务的节点需要等待没完成任务的节点，直到最后一个节点完成本地任务，所有节点统一进入下一轮的本地计算。

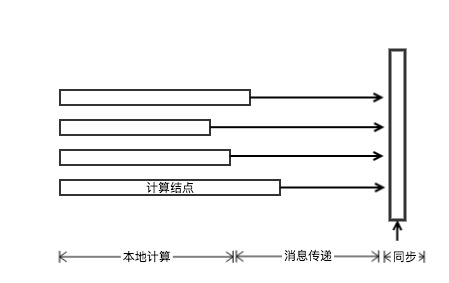


图2.2 超步内计算过程



图2.3 同步计算模型

从上图中我们可以清楚的看出，每轮计算中所有顶点只能使用上一轮迭代计算的更新消息，而通过计算得到的最新消息也只能通过同步在下一轮的计算中使用。因为全局同步的存在，使得整个图计算的调度十分简单，同时，每一轮的消息会在全局的同步下批量的传送，极大限度的利用了网络的流量。

为高效的处理大规模图数据的计算问题，在上述BSP计算模型的启发下，Google首次提出了以点为中心的计算框架Pregel[12]。在点中心的计算模型中，用户只需要从点的角度出发，重写一个与具体应用紧密相关的计算函数Compute，而不需要考虑数据的处理过程。在Compute函数内，涉及1) 处理上轮迭代的更新消息; 2) 本地计算逻辑，并更新本节点的值或节点状态信息; 3) 消息传递逻辑。 Pregel的具体执行过程如下: 首先，选定一个图分区策略[13-15]进行图分区操作，划分后的不同分区分配到不同的计算节点上，然后在每一个图节点上执行Compute函数，即用户自定义的超步处理逻辑，每一次的迭代后，通过全局同步操作保证所有计算节点统一进入下一轮的迭代计算，整个迭代过程直到没有任何消息需要同步或所有的图节点达到收敛状态时停止，输出结果。Pregel框架从容错性、一致性、同步控制等方面给出了可行的解决方案，整个计算模型简单、易用、可拓展性强。受Pregel框架的启发，相继涌现出很多基于BSP模型的大规模并行图计算系统，如GRAPE[18], GPS[16]、Giraph[17]、Hama、Giraph++、GiraphX、Pregel+、X-Pregel、Pregellix、Mizan、MOCgraph、Blogel、Nscale、SC-BSP、LFGraph等，是对Pregel开源实现的改进与扩展，在性能与功能上不断完善了基于BSP同步计算模型。

虽然BSP模型极大的简化的图计算任务，充分利用了网络流量，但是该模型仍有以下几方面限制：

(1) BSP模型具有木桶效应，因为全局同步的存在，每一轮迭代的计算时间取决于最慢的计算节点。因此，节点的计算能力不同或图数据的分区不合理都会导致计算节点间的负载不均衡，严重影响图计算效率。

(2) BSP模型每轮迭代只能使用上一轮迭代的更新信息，导致迭代次数较多，收敛速度较慢。

(3) 消息量大，导致通信代价昂贵。在分布式的图处理框架中，为了满足大规模图数据处理，通常需要将原始的图数据划分为多个子图，并把每个子图分配到不同的计算节点上，因为处于不同计算节点之间的顶点具有一定的关联关系，这就导致不同计算节点间需要进行消息传递以满足图计算的正确性。然而在实际的图数据中，尤其是社交网络等图应用，图中顶点的度数服从幂律分布，从而使得通信的代价成为影响系统性能的瓶颈之一。

(4) 某些需要协调图中邻接点的算法无法在BSP模型下实现，例如图着色问题[19],该问题意在使用最少的颜色为图着色，保证相邻的顶点颜色不同。这类问题对于贪心的解决方案在BSP模型下不会收敛，因为图节点试图获得非邻居节点的颜色值，只能通过两者的共同邻居节点来获得，因此同步操作的过程中可能会出现这两个节点颜色不断翻转的情况，本轮迭代具有相同颜色的邻接顶点在下轮的迭代仍会选取相同的颜色，从而导致无法收敛。

### 异步计算模型

现有的邮件重要性评价方法所使用的数据集多为私有商业数据集，公开的邮件数据集较少

### 混合计算模型

的

### 自适应计算模型

说

## 研究目标和内容

通过以上的国内外现状的分析和总结，我们可以看出国内外相关工作对于邮件重要性的评价方法忽略了邮件的上下文关系

## 论文组织结构

本文主要分为七个部分。主要的组织结构如下：

# 

# 相关理论与技术研究

由于本文提出的为基于会话网络的邮件重要性评价方法。

## 邮件重要性评价方法

邮件的重要性评价方法主要分为两类，。

### 基于社交网络的邮件重要性评价

基于社交网络的方法主要是通过邮件之间的通讯关系构建收发件人的社交网络

### 基于邮件特征的邮件重要性评价

由于基于邮件社交网络的方法忽略了邮件本身的结构信息

## 邮件会话网络的构建

### 基于元数据的会话网络构建

进行会话网络构建最简单的方法是使用邮件头中属性如IN-REPLY-TO和REFERENCE。

## 社交网络中节点重要性评价

在构建会话网络后。

### 基于网络拓扑的节点重要性评价

社交网络中，

### 基于交互信息的节点重要性评价

## 其他技术

## 本章小结

# 邮件重要性特征分析

由于之前的相关工作并没有验证哪些特征对于邮件重要性评价是有效的。

## 邮件特征

## 特征有效性实验

### 数据集及标注

### 分类器选择

# 邮件会话网络的构建

## 邮件会话网络的改进

## 邮件会话网络的生成

。

## 邮件会话网络的存储

## 实验评估

### 实验方案

# 基于会话网络的邮件重要性评价方法

本文的研究

## PageRank算法

PageRaEmailRank算法

在网页实验评估

### 实验方案

我们将第三本章小结

# EmailRank算法在实际系统中的应用

通过以上章

## 应用于邮件客户端

现有的邮件客户端如Gmail、Outook、Foxmail W3C会员影响力评价系统实现

### 需求分析

W3C是一个制定网络标准的全球会员组织总体架构设计

W3C会员影响力评价系统主要包括四大模块，系统模块实现

## 本章小结

本章我们介绍了提出的邮件重要性评价方法在实际系统的作用。

总结与展望

## 论文总结

在社交媒体和即时通讯飞速发展的如今，。

## 未来工作展望

本文主要研究了邮件重要性评价问题。

参考文献

1. The Radicati Group. Email-Market-2018-2022 [EB/OL]. https://www.radicati.com/wp /wp-content/uploads/2018/05/Email-Market-2018-2022-Executive-Summary.pdf, 2018-06-11/2018-11-20

攻读硕士学位期间取得的学术成果

**硕士期间发表的论文**

1. **Jiang Kun**, Hu chunming, Sun Jie, Shen Qi. Email importance evaluation in mailing list discussions [C]. International Conference on Web Information Systems Engineering. 2018. (CCF推荐会议列表C类，workshop，12页)
2. **Jiang Kun**, Hu chunming.：Characterizing and predicting important emails in mailing list discussions [C]. 北京航空航天大学第十五届研究生学术论坛. 2018.

**硕士期间申请的专利**

[1] 胡春明, **蒋坤**, 何潇安. 一种基于会话网络的邮件重要性评价方法[P]. CN：201811441956.X，2018-11-29.(专利学生第一发明人)

致 谢

在北京

。