统计分析说明书

组员：

SY1506404 孟翰

SY1506409 苏若

SY1506425 李璇

SY1506406 孙敏芳

“统计分析说明书”版本变更表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 版本号 | 文档名称 | 主要编制人 | 审核人 | 更新日期 | 备注 |
| 1.0 | 统计分析说明书 | 苏若、李璇 | 孟翰、孙敏芳 | 2016/05/06 |  |
| 2.0 | 统计分析说明书 | 苏若、李璇 | 孟翰、孙敏芳 | 2016/06/01 |  |
| 3.0 | 统计分析说明书 | 苏若、李璇 | 孟翰、孙敏芳 | 2016/06/13 |  |

一、目的

通过对软件复杂度、人员能力、实际工作量及工作日志因素进行综合性分析，使得对该课程设计中组内各成员的工作量以量化的形式输出，有助于之后本组对组内成员分工进行适当的调整，以发挥最高效的团队作用力。

1. 实验过程

2.1 工作复杂度分析

本节中，针对Apache Spark系统的复杂度及实施方案的复杂度建立了分析体系，主要分析结果如下：

2.1.1 Apache Spark系统的复杂度

1、系统剥离的复杂度

具体来看，本组选择的Apache Spark可概括为以下五方面：

a、一个围绕速度、易用性、复杂分析的大数据处理框架。

b、可统一管理不同性质（文本、图表）数据集。

c、与Hadoop的Mapreduce不同，Spark提供了分布式内存抽象（RDD），增加了新的操作(flatmap,filter,union,sample,join,groupByKey等）。

d、具有很高的容错性。

e、为迭代式数据处理提供更好的支持。每次迭代的数据可以保存在内存中，而不是写入文件。提高了并行计算效率。

正由于Spark体系结构的复杂与精细，其各组件各司其职又紧密交叉，加之我们对大数据领域知识接触较少，导致我们对该系统建立统筹的认识并能抽取出我们关注的重心是相对困难和复杂的。

2、环境搭建的复杂度

在实验过程中，本组采用了Linux下安装Spark集群环境的方法，因此，我们需要完成以下任务：

a、Linux系统的安装:采用Ubuntu 14.0及后续版本的.iso镜像文件实现安装，在此过程中涉及到系统的硬盘分区、系统基本环境配置等较多琐碎性工作。

b、Java JDK环境配置：采用Java JDK 1.7以上版本，并需要对环境变量进行配置。由于Linux系统的环境配置有多种方法，难免会出现配置错误。

c、Hadoop安装与配置：采用最新的Hadoop2.7.2版本，在进行环境变量配置前，需要完成一系列准备工作，如Hadoop用户的创建，SSH的设置以及Hadoop安装软件完整性的验证等。之后，需对Hadoop在Linux上的相关操作建立一定的了解，并完成Hadoop测试工作，为之后的集群环境的安装做好准备。

d、Spark 1.6.0安装与配置：采用最新的Spark 1.6.0版本。Spark配置可分为以下三种方式：单机版、伪分布式配置及集群安装配置，故本组在花费一定时间了解了这几种配置方式的区别后，结合待执行的任务特性，选择了集群的配置方式。在搭建过程中，主要涉及以下过程：

step1:选定一台机器作为 Master

step2:在 Master 节点上配置 hadoop 用户、安装 SSH server、安装 Java 环境

step3:在 Master 节点上安装 Hadoop，并完成配置

step4:在其他 Slave 节点上配置 hadoop 用户、安装 SSH server、安装 Java 环境

step5:将 Master 节点上的 /usr/local/hadoop 目录复制到其他 Slave 节点上

step6:在 Master 节点上开启 Hadoop

综上所看，该环境的搭建过程非常注重细节性，并且过程较多，且配置时间较长,较之其它开源系统来说，是更为复杂的一项工作。

2.1.2 实施方案的复杂度

在实验过程中，实验处理数据的收集也是一项耗时的任务，通过筛选，我们确定数据集“Movielens数据集”作为实验对象。

1、数据集概述

MovieLens 是历史最悠久的推荐系统。它由美国Minnesota 大学计算机科学与工程学院的GroupLens 项目组创办，是一个非商业性质的、以研究为目的的实验性站点。MovieLens 主要使用Collaborative Filtering 和Association Rules相结合的技术，向用户推荐他们感兴趣的电影。

MovieLens数据集中，用户对自己看过的电影进行评分，分值为1～5。MovieLens包括两个不同大小的库，适用于不同规模的算法．小规模的库是943个独立用户对1682部电影作的10000次评分的数据；大规模的库包括2万用户对1万部电影的1000万条评分数据，具有数据量大、数据权威、可验证机器学期的准确性以及具有实际应用价值等特点。

1. 实施方案复杂度

本项目选取movielens中的数据，采用spark-Mlib的recommendation协同过滤算法进行机器学习，希望能够通过用户的影评行为来对用户推荐他们感兴趣的电影，数据格式样例如下：

|  |  |  |
| --- | --- | --- |
| User | MovieId | RatingScore |
| 12487 | 45689 | 4.0 |

本项目将数据分为数据集与测试集（8:2），首先对数据进行去噪处理、数据提取等操作，利用Spark提供的机器学习中的推荐包对处理好的数据进行处理，对不同的参数进行测试以得到近似最优的准确率，用训练集训练后用测试集进行测试，根据测试准确率与spark 运行性能、调度策略来进行调优。

2.2 人员能力估计

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 组内成员 | 专业 | 系统了解度 | 编程能力 | 备注 |
| 孟翰 | 软件工程 | B | B | 组长 |
| 苏若 | 软件工程 | C | B |  |
| 李璇 | 软件工程 | C | B |  |
| 孙敏芳 | 软件工程 | C | B |  |

2.3 工作量统计与评估

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 任务编号 | 任务名称 | 任务描述 | 负责人 | 用时 | 完成情况 |
| 0 | 组队 |  | 苏若 | 2016.3.18~2016.3.18 | 完成 |
| 1 | 分配开源软件调研任务 | 调研各种开源软件 | 孟翰，苏若，李璇，孙敏芳 | 2016.3.19~2016.3.19 | 完成 |
| 2 | 查找资料 | 查找相关资料 | 孟翰，苏若，李璇，孙敏芳 | 2016.3.19~2016.3.23 | 完成 |
| 3 | 小组会议 | 决定选题 | 孟翰,苏若,李璇,孙敏芳 | 2016．3.23~2016.3.23 | 完成 |
| 4 | 制作PPT |  | 孟翰 |  | 完成 |
| 5 | 选题展示 |  | 孟翰 |  | 完成 |
| 6 | 问题清单整理 |  | 苏若 |  | 完成 |
| 7 | 项目计划书 |  | 孟翰,苏若,李璇,孙敏芳 | 2016.3.26~2016.3.26 | 完成 |
| 8 | 需求准备 | 学习RUCM、UML知识及工具使用 | 孟翰,苏若,李璇,孙敏芳 | 2016.3.27~2016.3.27 | 完成 |
| 9 | 系统建模 |  | 孟翰,苏若,李璇,孙敏芳 | 2016.3.28~2016.3.29 | 完成 |
| 10 | 需求分析报告 | 撰写需求分析 | 孟翰,苏若,李璇,孙敏芳 | 2016.3.30~2016.3.31 | 完成 |
| 11 | 需求分析报告展示 |  | 孟翰,苏若 | 2016.4.1~2016.4.1 | 完成 |
| 12 | 问题清单整理 |  | 李璇,孙敏芳 | 2016.4.1~2016.4.1 | 完成 |
| 13 | 小组会议 |  | 孟翰,苏若,李璇,孙敏芳 | 2016.4.2~2016.4.2 | 完成 |
| 14 | 改进版需求分析报告 |  | 孟翰,苏若,李璇,孙敏芳 | 2016.4.3~2016.4.7 | 完成 |
| 15 | 改进版需求报告评审 |  | 孟翰，苏若 | 2016.4.8~2016.4.8 | 完成 |
| 16 | 问题清单整理 |  | 李璇,孙敏芳 | 2016.4.9~2016.4.9 | 完成 |
| 17 | 小组会议 | 针对改进版需求评审，进行修正，并安排下一步工作 | 孟翰,苏若,李璇,孙敏芳 | 2016.4.10~2016.4.10 | 完成 |
| 18 | 最终版需求分析报告 | 撰写最终版需求分析 | 孟翰,苏若,李璇,孙敏芳 | 2016.4.11~2016.4.13 | 完成 |
| 19 | 软件项目进度控制与任务协调 | 使用project协调小组 | 孟翰,苏若,李璇,孙敏芳 | 2016.4.15~2016.4.20 | 完成 |
| 20 | 软件变更控制与版本管理 | 使用github变更和管理 | 孟翰,苏若,李璇,孙敏芳 | 2016.4.21~2016.4.21 | 完成 |
| 21 | 研究确定变更技术方案目标 |  | 孟翰,苏若,李璇,孙敏芳 | 2016.4.21~2016.4.21 | 完成 |
| 22 | 小组会议 | 讨论确定新的方案 | 孟翰,苏若,李璇,孙敏芳 | 2016.4.22~2016.4.22 | 完成 |
| 23 | 新的方案目标实现 |  | 孟翰,苏若,李璇,孙敏芳 | 2016.4.23~2016.5.5 | 完成 |
| 24 |  |  |  |  |  |

//