中国科学技术大学软件学院

软件工程实验项目环节

开题报告

项 目 名 称：大数据环境下集成 R 语言的数据挖掘平台

成 员 名 单： 张杰 张静 邱星 李大学 叶庆

导　　师： 周东仿

工 程 领 域： 软件工程

研 究 方 向： 数据挖掘

开 题 时 间：

中国科学技术大学软件学院

填表日期：　　年　月　日

一、 简况

<table>
<tr><td rowspan="2">名称</td><td>中文</td><td colspan="3">大数据环境下集成 R 语言的数据挖掘平台</td></tr>
<tr><td>英文</td><td colspan="3">The Data Mining Platform Integrated With R Language Under Big Data Environment</td></tr>
<tr><td rowspan="6">项目组成员名单</td><td>姓名</td><td>学号</td><td>项目中的分工</td><td>签 章</td></tr>
<tr><td>张杰</td><td>SA13226427</td><td>数据挖掘算法实现</td><td></td></tr>
<tr><td>邱星</td><td>SA13226267</td><td>Web 前端开发</td><td></td></tr>
<tr><td>张静</td><td>SA13226428</td><td>Hadoop 环境搭建和性能优化</td><td></td></tr>
<tr><td>李大学</td><td>SA13226201</td><td>数据读取和预处理</td><td></td></tr>
<tr><td>叶庆</td><td>SG13225145</td><td>后台连接 R 语言，数据可视化显示</td><td></td></tr>
<tr><td>中文摘要</td><td colspan="4">
摘要

随着信息化的推进，企业产生了大量业务数据，其中蕴藏着大量未知的、潜在的信息。数据挖掘是一种新的商业信息处理技术，通过对商业数据库中的大量业务数据进行抽取、转换、分析和其他模型化处理，可提取辅助商业决策的关键性信息。

开源软件 R 是当今相当流行的数据分析、统计制图语言，提供了丰富的分析模块和实用工具，在业界已得到广泛应用。但是，最终用户并非都是数据分析专家，难以快速掌握和灵活运用 R 语言。为了充分发挥 R 语言的价值，为用户提供功能强大的分析工具，设计一个集成了 R 语言功能的、易用的数据挖掘平台意义重大。

本项目设计一个能处理大数据的、以 R 语言作为数据分析引擎的“大数据环境下集成 R 语言的数据挖掘平台”，以解决企业在数据挖掘方面日益增长的需求。
</td></tr>
</table>

Abstract

With the advance of informatization, enterprises have produced a large volume business data, which contain a large number of unknown, potential information. Data mining, which is a new business information processing technology, could extract the key information for supporting the business decision through the process of extraction, transformation, analysis, modeling based on a large amount of business data stored in the commercial database.

R is an open source software which provide a language for data analysis and statistical mapping. It provides a rich module for analysis and a branch of utilities so it has been a widely used in the industry. However, the end user is not all data analysis expert, it is difficult for end users to grasp and use the R flexibly. In order to give full play to the value of R language and provide a set of powerful utilities for user, it is essential to designed user-friendly data mining platform integrated with R language.

This project is to design a "Data Mining Platform in big data environment with R language integrated "that can handle large data, using R language as a data analysis engine , in order to solve the increasing demand of enterprise in the aspect of data mining.

英文摘要

主题词		
	中文	数据挖掘 分析与预测 R 语言
	英文	Data mining/Analysis and forecast/R language

二、选题依据

1．阐述该选题的研究意义，或工程设计的价值和意义，国内外概况和发展趋势，选题的先进性和实用性，技术难度及工作量。

(1) 研究意义

随着信息科技的进步以及电子化时代的到来，人们以更快捷、更容易、更廉价的方式获取和存储数据，使得数据及信息量以指数方式增长。据粗略估计，一个中等规模企业每天要产生 100MB 以上的商业数据。而电信、银行、大型零售业每天产生的数据量以 TB 来计算。快速增长的海量数据收集、存放在大型数据库中，如果没有强有力的工具，理解它们已经远远超出了人的能力范围，收集在大型数据库中的海量而杂乱数据变成了“数据垃圾”、“数据坟墓”。高维海量的数据增加了传统统计分析方法的难度，这样，对大型数据的处理和分析的需求显得越来越迫切。如何才能不被信息的汪洋大海所淹没，从中及时发现有用的知识，提高信息利用率呢?要想使数据真正成为一个公司的资源，只有充分利用它为公司自身的业务决策和战略发展服务才行。因此，面对“人们被数据淹没，人们却饥饿于知识”的挑战，从数据库中发现知识(Knowledge Discovery in Databases)及其核心技术——数据挖掘(Data Mining)便应运而生，并得以蓬勃发展，越来越显示出其强大的生命力。所以基于知识发现及决策支持的需要，构建一个数据挖掘平台，并通过这个平台帮助人们从中发现有用信息及知识是很有意义的。

数据挖掘的工具很多，但绝大部分产品的价格都是相对比较昂贵，开源软件 R 是当今相当流行的数据分析、统计制图语言，提供了丰富的分析模块和实用工具，在业界已得到广泛应用。但是，最终用户并非都是数据分析专家，难以快速掌握和灵活运用 R 语言。为了充分发挥 R 语言的价值，为用户提供功能强大的分析工具，设计一个集成了 R 语言功能的、易用的数据挖掘平台意义重大。

Hadoop 是一个分布式系统的基础架构，且 HDFS 可以部署在低廉的硬件上。我们只用比较低的成本就可以充分利用集群的威力高速运算和存储，进而利用开源软件 R 语言进行数据分析，最终运用数据挖掘中的分类、聚类、关联及预测中的一种或多种分析方法来进行分析及预测，从而发现商机。

(2) 国内外发展趋势

从数据库中发现知识(KDD)一词首次出现在 1989 年举行的第十一届国际联合人工智能学术会议上。随后在 1991 年、1993 年和 1994 年都举行 KDD 专题讨论会，汇集来自各个领域的研究人员和应用开发者，集中讨论数据统计、海量数据分析算法、知识表示、知识运用等问题。随着参与人员的不断增多，KDD 国际会议发展成为年会。到目前为止，由美国人工智能协会主办的 KDD 国际研讨会已经召开了 8 次，规模由原来的专题讨论会发展到国际学术大会，研究重点也逐渐从发现方法转向系统应用，注重多种发现策略和技术的集成，以及多种学科之间的相互渗透。1999 年，亚太地区在北京召开的第三届 PAKDD 会议收到 158 篇论文，空前热烈。IEEE 的 KnowledgeandDataEngineering 会刊率先在 1993 年出版了 KDD 技术专刊。并行计算、计算机网络和信息工程等其他领域的国际学会、学刊也把数据挖掘和知识发现列为专题和专刊讨论，甚至到了脍炙人口的程度。数据挖掘在 1995 年召开了第一届知识发现与数据挖掘国际学术会议。该会议是由 1989 年至 1994 年举行的四次数据库中知识发现国际研讨会发展来的。数据挖掘研究界于 1998 年建起了一个新的学术组织 ACM-SIGKDD，即 ACM 下的数据库中知识发现专业组(Special Interested Groupon Knowledge Discovery in Database).1999 年 ACM-SIGKDD 组织了第五届知识发现与数据挖掘国际学术会议(KDD'99)。专题杂志 DataMining and Knowledge Discovery 自 1997 年起有 Kluwers 出版社出版。ACM-SIGKDD 还出版了一种季刊电子通信 SIGKDDExplorations。还有一些其他国际或地区性的数据挖掘会议如“知识发现与数据挖掘太平洋亚洲会议”(PAKDD),“数据库与知识发现原理与实践欧洲会议”(PKADD)和“数据仓库与知识发现国际会议(DaWaK)涉及数据挖掘的研究成果已在许多数据库国际会议论文集发表，包括"ACM-SIGMOD 数据管理国际会议”(SIGMOD),“超大型数据库国际会议”(VLDB),"ACM-SIGMOD-SIGART 数据库原理研讨会”(PODS),“数据工程国际会议”(ICDE),“扩展数据库技术国际会议”(EDBT),“数据库理论国际会议”(ICDT)“信息与知识管理国际会议”(CIKM),“数据库与专家系统应用国际会议”(DEXA)和“数据库系统高级应用国际会议”(DASFAA)数据挖掘的研究也发表在主要数据库杂志上，包括《IEEE 知识与数据工程汇刊》(TKDE),(ACM 数据库系统汇刊》(TODS),(ACM 杂志》(JACM),《信息系统》，(VLDA 杂志》,《数据与知识工程》，和《智能信息系统国际杂志》(JIIS)。

此外，在 Internet 上还有不少 KDD 电子出版物，其中以半月刊 KnowledgeDiscoveryNuggets 最为权威。目前，世界上比较 有影响的典型数据挖掘系统有:SAS 公司的 EnterpriseMiner,IBM 公司的 IntelligentMiner,SGI 公 司的 SetMiner,SPSS 公司的 Clementine,Sybase 公司的 WarehouseStudio,RuleQuestResearch 公司 的 See5、还有 CoverStory, EXPLORA, KnowledgeDiscoveryWorkbench,Miner,Quest 等。

国内从事数据挖掘研究的人员主要在大学，也有部分在研究所或公司。所涉及的研究领域很多，一般集中于学习算法的研究、数据挖掘的实际应用以及有关数据挖掘理论方面的研究。目前进行的大多数研究项目是由政府资助进行的，如国家自然科学基金、863 计划、“九五”计划等，但还没有关于国内数据挖掘产品的报道。与国外相比，国内对数据挖掘的研究稍晚，没有形成整体力量。1993 年国家自然科学基金首次支持对该领域的研究项目。目前，国内的许多科研单位和高等院校竞相开展知识发现的基础理论及其应用研究，如清华大学、中科院计算技术研究所、空军第三研究所、海军装备论证中心等。北京系统工程研究所对模糊方法在知识发现中的应用进行了较深入的研究，北京大学也在开展对数据立方体代数的研究；华中理工大学、复旦大学、浙江大学、中国科技大学、中科院数学研究所、吉林大学等单位开展了对关联规则挖掘算法的优化和改造；南京大学、四川联合大学和上海交通大学等单位探讨、研究了非结构化数据的知识发现以及 Web 数据挖掘。

(3) 实践工作量及技术难度

本实践的工作量主要集中以下几个方面：

1) JavaEE 系统平台的构建。
2) 团队开发平台（Git）的搭建
3) 数据挖掘相关知识的学习。
4) R 语言进行数据挖掘的算法。
5) Hadoop 分布式文件系统的搭建及配置。
6) 处理多种数据源。
7) web 前端开发。
8) 大数据处理的性能优化。

主要的技术难点在于团队合作开发，新技术的学习与使用以及如何将项目整合。另外非结构化数据的处理也是本项目的重点与难点。

2 .参考文献

[1] 数据挖掘研究现状及发展趋势[B]. 王惠中,彭安群. 工矿自动化 2011 年 2 月第 2 期.

[2] 基于 JVM 的 R 语言海量数据统计集成框架研究.曹杰. 华中科技大学 硕士学位论文

[3] 胡侃, 夏绍玮. 基于大型数据仓库的数据采掘: 研究综述[J]. 软件学报, 1998, 9(1): 53-63.

[4] 李军华. 云计算及若干数据挖掘算法的 MapReduce 化研究[D]. 成都: 电子科技大学硕士论 文, 2010.

[5] 陈娜. 数据挖掘技术的研究现状及发展方向[J]. 电脑与信息技术, 2006, 2(1): 46-49.

[6] 李寒. Hadoop 关联数据挖掘 aprior 的研究与实现[D].桂林：桂林电子科技大学,2012.

[7] Tom White.Hadoop 权威指南[M].北京：清华大学出版社.2011.

[8] Robert I.Kabacoff.R 语言实战[M].北京：人民邮电出版社.2013.

[9] 方匡南. 基于数据挖掘的分类和聚类算法研究及 R 语言实现[D].广州：暨南大学，2007.

[10] Revolution Analytics.(2013) RHadoop project on Github website.[Online]. Available:https://github.com/RevolutionAnalytics.

[11] HAN Jiawei, KAMBER M. 数据挖掘: 概念与技术[M] . 范明, 孟小峰, 译. 北京: 机械工业出版社, 2001.

[12] 纪俊.一种基于云计算的数据挖掘平台架构设计与实现[C].青岛大学，2009.6

[13] 陕粉丽. 数据挖掘技术的研究现状及应用[J] . 现代企业教育, 2008(6): 101-102.

三、课题内容及具体方案

1．课题内容

(1) 开发 web 前端界面，具备良好的用户体验。

(2) 开发 web 服务器端，调用挖掘的具体过程。

(3) 针对不同需求进行 R 语言数据挖掘算法开发。

(4) Hadoop 分布式处理系统搭建。

2．系统需求分析

(1) 用户使用图形化界面进行操作。用户可以设置参数来源，选择分析方法，设置分析参数，建立分析流程，不用编写 R 代码就能够进行数据分析，得出结果。

(2)系统提供分类、聚类、关联规则、预测等数据挖掘分析方法，并且对这些方法提供参数设计界面，允许用户通过调用参数，优化分析结果。

(3) 系统能够处理多种来源数据。

(4) 系统能够处理结构化、半结构化以及分结构化数据。

(5) 系统能够形象化的分析结果展现。能够以表和图的形式显示分析结果，并将结果导出成 xsl 和 pdf 文件。

(6) 在处理大数据时不出现内存溢出。

(7) 具有较好的响应速度，处理 100 万行记录以下的数据时，不应超过 10 秒（能够处理 TB 级以上的结构化数据将更好）。

(8) 具有较好的可扩展性，包括对系统功能的扩展（比如增加分析方法不影响现有系统）和计算资源的扩展（如果使用集群提升性能，新加入的计算机不需要对现有架构作大的调整）等。

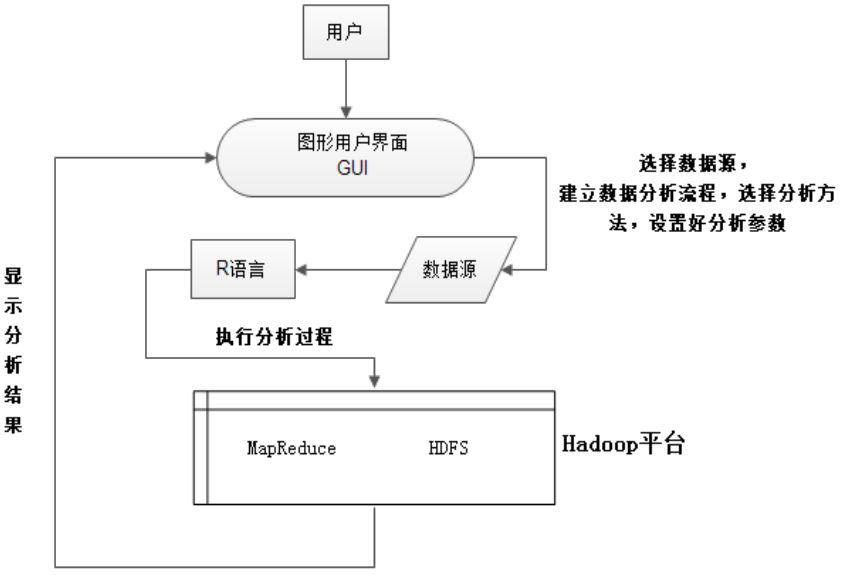
系统需求分析由下表列出：

表 3.1 系统需求分析表

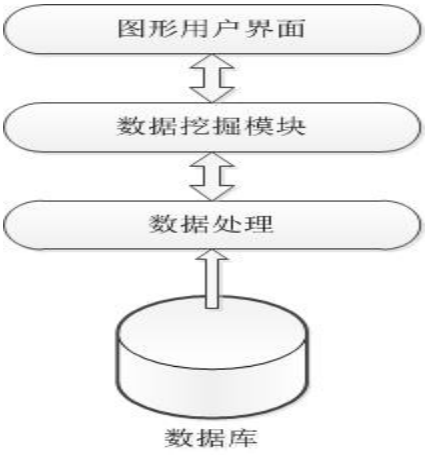
功能性需求	**输入**	**用户操作模式**	**Web 界面**
		挖掘方法输入	**分类**
			聚类
			预测
		数据库输入	**Oracle**
			MySql
		文件输入	**Csv 及 xls 文件**
	输出	**Web 页面输出**	**图或表**
		文件输出	**pdf 文件,xls 文件**
非功能性需求	**稳健性**	**不出现内存溢出**	
	性能	**100 万记录行在 10S 以内**	
	可扩展性	**系统功能及计算资源的扩展**	

3. 系统设计

系统概要设计：



系统总体结构：



4. 拟采用的开发方法、环境，测试方案等

开发方法：面向对象方法

编程语言：JAVA

IDE 工具：eclipse

数据库：Oracle, MySql

操作系统：Linux(Ubuntu 13.04)

搭建平台：Hadoop（软件开发环境），Git（版本管理工具）

硬件开发环境如下：

电脑：五台联想昭阳 E46L 笔记本

CPU：Intel 赛扬双核 T3100

内存：2GB DDR3 1066MHz

硬盘：320GB

测试方案：三轮测试。即单元测试，系统测试及用户验收测试。

5．技术难度及特色分析

技术难度如下：

- 团队开发平台（Git）的搭建
- Hadoop 开发环境搭建
- R 语言的学习与使用
- 数据挖掘分析方法（4 种）的学习与使用
- 处理多种数据源
- 良好的图形界面
- 非结构化数据的处理

形象化的分析结果展现，数据预测，支持多种数据源的导入，高响应度，可扩展性及同时处理结构化与非结构化的数据都将是本项目的特色。其中数据预测及非结构化的数据的处理将是本项目的重点难点以及最大的特色。

四、工作进度的大致安排

应包括文献调研，理论分析，数值计算，理论分析，软硬件设计，仪器设备的研制和调试，撰写结题报告、技术论文等，要给出各个阶段的成果形式。

2013.11 月——开题报告，开题答辩

2013.12 月——完成详细设计、搭建开发环境（Git 及 Hadoop）。

2014.1 月——初步完成每个模块的实现

2014.3 月——组员分别完成每个模块的实现，调试

2014.4 月——项目进行整合，调试

2014.5 月——完成项目文档，撰写结题报告、技术论文

2014.6 月——答辩

预期成果及特色

- ✓ 良好的图形化界面，用户可以设置数据来源，选择分析方法，设置分析参数
- ✓ 优秀的用户体验
- ✓ 系统提供主流的数据挖掘算法
- ✓ 能够处理多种来源数据
- ✓ 能够以表和图的形式显示分析结果，并将结果导出成 xls 和 pdf 文件
- ✓ 具有较好的响应速度，处理 100 万行记录以下的数据时，不超过 10 秒
- ✓ 具有较好的可扩展性

导师意见（对选题和工作过程及成果进行说明，并给出成绩。）   导师签名： 年　　月　　日
答辩小组意见   答辩组长签名： 年　　月　　日