**调研报告**

**课题：基于Spark的大数据平台的实现**

**组员：SA18225084 甘朔**

**SA18225356 王翰青**

**SA18225455 游奕航**

**SA18225515 钟伟**

1. **前沿**

1.历史背景：

“大数据”(Big Date)是 2008年以来一个新兴的概念，自其产生之初便备受关注，其衍生的技术也随之发展。大数据平台和数据挖掘便是大数据处理中重要的技术组成。大数据平台可以实现数据的分布式处理，提高数据的处理效率，降低时延，保证数据的有效性。

中国计算机学会大数据委员会在2012年12月召开的大数据学术会议中预测，201年大数据最令人瞩目的六大应用领域为互联网、金融、健康医疗、企业、政府管理及安全。金融大数据作为第二大最令人瞩目的应用，也是股市预测、金融分析等研究的关键问题。IBM高级工程师在介绍Mo Sphere Streams行业用例时，预测金融服务领域将是大数据分析技术的第一应用领域。

1. 大致发展：

云计算、物联网、社交网络等新兴信息技术的快速发展，促使全球数据量急剧增加，推动人类社会迈入大数据时代。企业界和学术界将大数据的5个“V”归纳为数据量（Volume）、数据类别（Variety）、数据价值（Value）、处理速度（Velocity）以及数据真实性（Veracity）。大数据呈现出多种鲜明特征，数据量以倍增趋势增加；数据类型多样，半结构化和非结构化数据所占比例迅速增加；数据量增加到一定规模，数据隐含的价值也随之增大，但随着时间推移，数据价值的有效性随时间迅速减少，对数据计算能力以及数据真实性的要求不断提。特别是在金融行业，股票信息的增加致使数据量的增长尤为庞大，从庞大的数据量中分析出我们想要的信息，甚至让机器帮我们分析，成为现在越来越热门的趋势。

1. 现实状况和主要成绩：

随着各个行业逐渐流入互联网市场，行业在数据流量上的增长也日渐明显，但是行业中大数据的挖掘工作还需要进一步加强，例如在分析大数据方面，我们还有很多难题需要解决，而如何高效智能的对大数据挖掘算法进行分析并建立一个能够分析海量大数据的平台，也是当下急需处理的难点，处于这个数据决定一切的时代，建立好的平台高效的处理数据非常重要，Spark 能够很好地完成这个使命，Spark 能够在大数据分析可靠性上提供保障，利用分布式计算框架提供更多分析大数据的依据，并且支持更多更复杂的查找，在高速运行的数据处理中提供更加准确的服务。然而影响股票信息的因素众多，包括人为的因素等一系列不可预测因素。

在股票市场上，一般采用两类股票投资分析方法:基本分析法和技术分析法。其中基本分析法是通过对影响股票市场供求关系的基本因素进行分析，以此来确定股票的真正价值，判断股票市场的未来走势，提供投资者选择股票交易的依据。而技术分析法则是一种完全根据股票市场行情变化而加以分析的方法，它通过对历史资料(如股票的成交量和成交价格)进行分析，判断整个股票市场或是个别单只股票价格未来的变化趋势，给投资者提供交易股票的信号。这其中建立在统计学基础之.上常见的技术分析方法，有移动平均线法、点数图法、K线图法等。对于受政治、经济、心理、国际市场等多种因素影响的复杂股票市场，使用传统的技术分析工具进行股票买卖决策难度较大，大多数投资者应用的结果并不理想。

1. **研究目的与意义**

股票作为股份公司发行的所有权凭证，是一种股份公司为筹集资金而发行的一种有价证券。企业或个人通过积累货币执行股票的买入买出，借此实现收益。然而股票市场交易价格往往一日数十变，价涨获利，价跌亏损，一旦股价下跌，投资者就可能受到损失。本课题希望通过使用spark大数据平台对实时股市行情进行分析，结合历史数据和特征，运用如机器学习，人工智能等热点新兴技术对股票发展趋势进行预测，降低投资人受损的可能性。

1. **国内外研究现状**

文献[1]是最早使用机器学习和支持财务决策的文本挖掘技术之一。文中釆用KNN、反向神经网络和朴素贝叶斯三种不同的机器学习方法，通过预先定义的金融词典，利用关键词与金融市场的表现产生概率规则，验证了规则技术对于股票价格走势预测的有效性，并发现预测结果的准确率明显高于随机预测。

文献[2,3]发现大多数股票预测研究，只是对股票历史数据进行简单的获取和处理，忽略了现实世界事件的影响。基于此观点，它采用历史数据与新闻相结合的预测方法，通过实验证实了现实事件对股票价格有较大影响。

文献[4]釆用朴素贝叶斯算法对三万个人工情感标注的评论作为训练样本，1994年到2007年的公司前瞻性陈述的内容作为测试数据，验证了公司管理人员讨论与分析的积极性与未来的股票价格的波动与收益是相关的。

文献[5]研究Yahoo! Finance和Raging Bulling网站上超过150万条消息对45家公司的道琼工业平均指数和道琼指数的影响，发现股评消息可以预测股票市场的波动，并且存在分歧的消息与交易量增加之间存在明显相关性。

文献[6]釆用支持向量机回归算法对华尔街日报流行专栏的内容和股票价格进行定量分析，发现媒体悲观预测股票价格会导致股票价格下跌，以及异常高或低的悲观预测会对股票交易量产生较大影响。

文献[7]从股票留言板提取投资者情绪，根据不同的分类算法对其进行情感分析。实验结果发现，贝叶斯分类器使用时间序列和消息相结合，改善了所得情绪指数的质量，并验证了投资者情绪对于股票走势的影响。

**四、课题研究内容**

根据项目要求，我们将会在建立了大数据平台的基础上实现一个数据分析系统，并用于股票数据的分析。以此为依据，研究内容可分为两个部分：基础技术和领域方向。基础技术描述了计算机科学技术的相关内容，是领域方向的支撑。领域方向则描述了股票数据分析的一些方面，具有更强的针对性。

1. 基础技术：

1.集群部署

Apache Spark是一个开源集群运算框架，最初是由加州大学柏克莱分校AMPLab所开发。相对于Hadoop的MapReduce会在运行完工作后将中介数据存放到磁盘中，Spark使用了存储器内运算技术，能在数据尚未写入硬盘时即在存储器内分析运算。Spark在存储器内运行程序的运算速度能做到比Hadoop MapReduce的运算速度快上100倍，即便是运行程序于硬盘时，Spark也能快上10倍速度。Spark允许用户将数据加载至集群存储器，并多次对其进行查询，非常适合用于机器学习算法。

使用Spark需要搭配集群管理员和分布式存储系统。Spark支持独立模式（本地Spark集群）、Hadoop YARN或Apache Mesos的集群管理。[在分布式存储方面，Spark可以和HDFS、 Cassandra 、OpenStack Swift和Amazon S3等接口搭载。 Spark也支持伪分布式（pseudo-distributed）本地模式，不过通常只用于开发或测试时以本机文件系统取代分布式存储系统。

Spark应用程序作为集群上的独立进程集运行，由SparkContext 主程序中的对象（称为驱动程序）协调。

具体来说，要在集群上运行，SparkContext可以连接到几种类型的集群管理器 （Spark自己的独立集群管理器，Mesos或YARN），它们跨应用程序分配资源。连接后，Spark会在集群中的节点上获取执行程序，这些节点是为您的应用程序运行计算和存储数据的进程。接下来，它将您的应用程序代码（由传递给SparkContext的JAR或Python文件定义）发送给执行程序。最后，SparkContext将任务发送给执行程序以运行。



图4-1

（DriverProgram 就是程序员自己设计的Spark程序，在Spark程序中必须定义SparkContext，它是开发Spark应用程序的入口。SparkContext通过Cluster Manager管理整个集群，集群中包含多个Work Node，在每个WorkNode中都有Executor负责执行任务。）

2.Scrapy爬虫技术

Scrapy是一个为了爬取网站数据，提取结构性数据而编写的应用框架。 其可以应用在数据挖掘，信息处理或存储历史数据等一系列的程序中。其最初是为了页面抓取 所设计的， 也可以应用在获取API所返回的数据 或者通用的网络爬虫。Scrapy用途广泛，可以用于数据挖掘、监测和自动化测试。此外，Scrapy 还使用了 Twisted异步网络库来处理网络通讯。



图4-2

3.Spark MLlib 机器学习

MLlib是Spark的机器学习库，旨在简化机器学习的工程实践工作，并方便扩展到更大规模。MLlib由一些通用的学习算法和工具组成，包括分类、回归、聚类、协同过滤、降维等，同时还包括底层的优化原语和高层的管道API。具体来说，其主要包括以下几方面的内容：

算法工具：常用的学习算法，如分类、回归、聚类和协同过滤；

特征化公交：特征提取、转化、降维，和选择公交；

管道(Pipeline)：用于构建、评估和调整机器学习管道的工具;

持久性：保存和加载算法，模型和管道;

实用工具：线性代数，统计，数据处理等工具。

使用 ML Pipeline API可以很方便的把数据处理，特征转换，正则化，以及多个机器学习算法联合起来，构建一个单一完整的机器学习流水线。这种方式给我们提供了更灵活的方法，更符合机器学习过程的特点，也更容易从其他语言迁移。Spark官方推荐使用spark.ml。如果新的算法能够适用于机器学习管道的概念，就应该将其放到spark.ml包中，如：特征提取器和转换器。

Spark在机器学习方面的发展非常快，目前已经支持了主流的统计和机器学习算法。纵观所有基于分布式架构的开源机器学习库，MLlib可以算是计算效率最高的。MLlib目前支持4种常见的机器学习问题: 分类、回归、聚类和协同过滤。

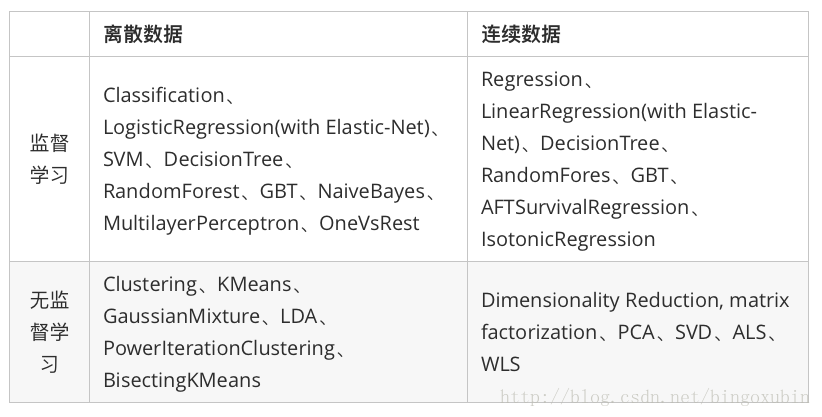


图4-3

1. 核心算法

通过对于股票预测的相关知识的调研补充，对具体的预测数学模型进行分析，并综合其优缺点总结为表4\_1：

表 4\_1：

|  |  |  |  |
| --- | --- | --- | --- |
| 预测模型名称 | 适用范围 | 优点 | 缺点 |
| 线性回归模型 | 线性回归是经济学的主要实证工具。例如，它是用来预测消费支出，固定投资支出，存货投资，一国出口产品的购买，进口支出，要求持有流动性资产，劳动力需求、劳动力供给。 | 建模速度快，不需要很复杂的计算，在数据量大的情况下依然运行速度很快。  可以根据系数给出每个变量的理解和解释 | 对异常值很敏感，结果误差较大。 |
| 时间序列预测模型 | 根据客观事物发展的这种连续规律性，运用过去的历史数据，通过统计分析，进一步推测市场未来的发展趋势。时间序列在时间序列分析预测法处于核心位置。 | 一般用ARMA模型拟合时间序列，预测该时间序列未来值。Daniel检验平稳性。自动回归AR(Auto regressive)和移动平均MA(Moving Average)预测模型。 | 当遇到外界发生较大变化，往往会有较大偏差，时间序列预测法对于中短期预测的效果要比长期预测的效果好。 |
| 马尔可夫预测模型 | 适用于随机现象的数学模型（即在已知现情况的条件下，系统未来时刻的情况只与现在有关，而与过去的历史无直接关系） | 研究一个商店的未来某一时刻的销售额，当现在时刻的累计销售额已知。 | 不适宜用于系统中长期预测 |
| BP神经网络 | 1)函数逼近：用输入向量和相应的输出向量训练一个网络逼近一个函数。  2)模式识别：用一个待定的输出向量将它与输入向量联系起来。  3)分类：把输入向量所定义的合适方式进行分类。  4)数据压缩：减少输出向量维数以便于传输或存储。 | 1) 非线性映射能力  2) 自学习和自适应能力  3) 泛化能力  4) 容错能力 | 1) 局部极小化问题  2) BP 神经网络算法的收敛速度慢  3) BP 神经网络结构选择不一  4) 应用实例与网络规模的矛盾问题  5) BP神经网络预测能力和训练能力的矛盾问题  6） BP神经网络样本依赖性问题 |
| 支持向量机（SVM） | 支持向量回归主要用于处理函数回归的问题，目前在系统识别、系统预测中的应用已取得比较好的效果。 | 泛化错误率低，计算开销不大，结果易于解释。  支持向量机就克服了神经网络收敛速度慢和局部极小点等缺陷。 | 对参数调节和核函数的选择敏感，原始分类器不加修改仅适用于处理二类问题。 |

回归模型介绍：

回归模型的基本思想在于根据历史数据的变化规律，寻找自变量与因变量之间的回归方程式，确定模型参数，据此预测。回归问题可分为一元和多元回归、线性和非线性回归。

3.3.1线性回归模型

线性回归是指全部由线性变量组成的回归模型。在实际生活中，由于单一变量不能满足描述输出变量的关系，需要使用多个变量来表示与输出的关系，即所谓多变量线性回归，形如：



其中a为系数，x为变量，b为偏置。但因为这个函数只有线性关系，所以只适用于建模线性可分数据。

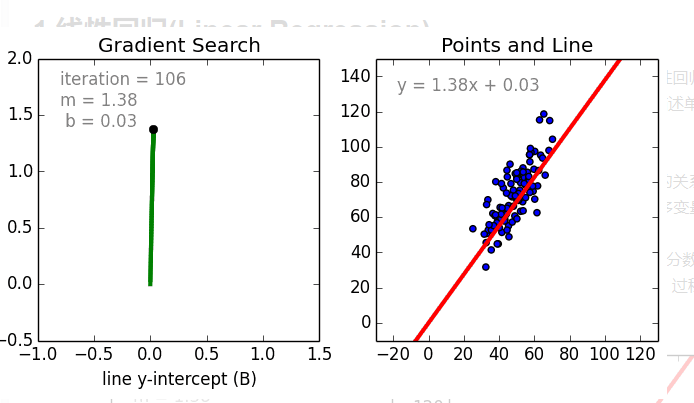


图4-4

3.3.2多项式回归模型

面对非线性可分数据时使用多项式回归，这种回归中，我们通过一条曲线来拟合数据点，形如：



多项式回归能够拟合非线性可分的数据，更加灵活的处理复杂的关系。它可以完全控制要素变量，配合一些数据的先验知识才能得到最佳指数。若指数选择不当便会造成过拟合现象。

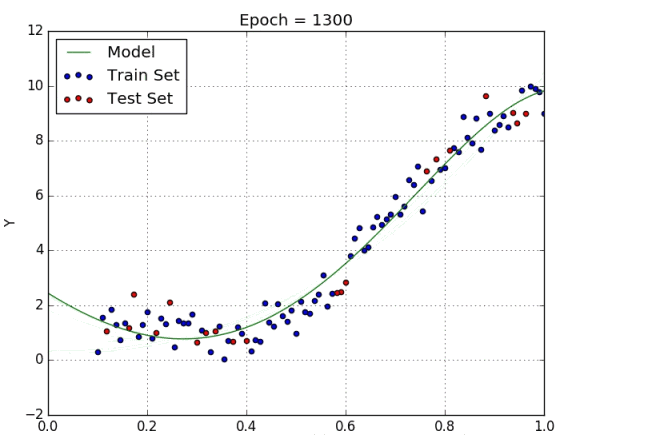


图4-5

3.3.2时间序列预测模型

时间序列是指将同一统计指标的数值按其先后发生的时间顺序排列而成的数列。其使用目的是根据已有的历史数据对未来进行预测。常用的四种时间序列模型分别为：自回归型AR（p），移动平均模型MA（q），自回归移动平均模型ARMA（p，q），自回归差分移动平均模型ARIMA（p，d，q），前三种模型可视为ARIMA（p，d，q）的特殊情况。

3.3.2.1 ARIMA（p，d，q）模型

ARIMA模型含义是单积自回归移动平均过程，其含义为：假设一个随机过程含有d个单位根，其经过d次差分后可以变换为一个平稳的自回归移动平均过程，则称该随机过程为单积自回归移动平均过程。

（1）平稳随机过程

若一个随机过程m阶以下的矩的取值全部与时间无关，则称该过程为m阶平稳过程。通常我们使用一阶平稳过程，即随机过程xt的均值mt不随时间变化

（2）自回归过程

如果一个剔除均值和确定性成分的线性过程可以表达为



其中，是自回归参数，是白噪声过程，则称为p阶自回归过程，用AR(p)表示。

（3）移动平均过程

如果一个剔除均值和确定性成分的随机过程可以用下式表达为：



其中，是自回归参数，是白噪声过程，则称上式为q阶移动平均过程，记为MA(q)。之所以成为移动平均，是因为是由q+1个及其滞后项加权构成的。这个“移动平均”需要和移动平均算子做区分。

（5）自回归移动平均过程

如果一个剔除均值和确定成分的线性随机过程由自回归和移动平均两部分共同构成，则称其为自回归移动平均过程，记为ARMA(p,q)，表示如下：



（6）差分：时间序列变量的本期值与其滞后值相减的运算称为差分。举例如，一阶差分：



ARIMA模型首先通过n次差分后，判断该随机过程是否平稳；随后找到合适的d后，将转化为平稳的随机过程，将构建为自回归移动平均过程。

3.3.3马尔可夫预测模型

马尔可夫过程是一类随机过程。它的原始模型马尔可夫链，由俄国数学家A.A.马尔可夫于1907年提出。该过程具有如下特性：在已知目前状态（现在）的条件下，它未来的演变（将来）不依赖于它以往的演变 (过去 )。例如森林中动物头数的变化构成——马尔可夫过程。在现实世界中，有很多过程都是马尔可夫过程，如液体中微粒所作的布朗运动、传染病受感染的人数、车站的候车人数等，都可视为马尔可夫过程。

一个马尔科夫过程就是指过程中的每个状态的转移只依赖于之前的 n个状态，这个过程被称为1个 n阶的模型，其中 n是影响转移状态的数目。最简单的马尔科夫过程就是一阶过程，每一个状态的转移只依赖于其之前的那一个状态。

马尔可夫链是指数学中具有马尔可夫性质的离散事件随机过程。该过程中，在给定当前知识或信息的情况下，过去对于预测将来是无关的。在马尔可夫链的每一步，系统根据概率分布，可以从一个状态变到另一个状态，也可以保持当前状态。状态的改变叫做转移，与不同的状态改变相关的概率叫做转移概率。 马尔可夫链描述了一种状态序列，其每个状态值取决于前面有限个状态。马尔可夫链是具有马尔可夫性质的随机变量的一个数列。这些变量的范围，即它们所有可能取值的集合，被称为“状态空间”，而的值则是在时间n的状态。

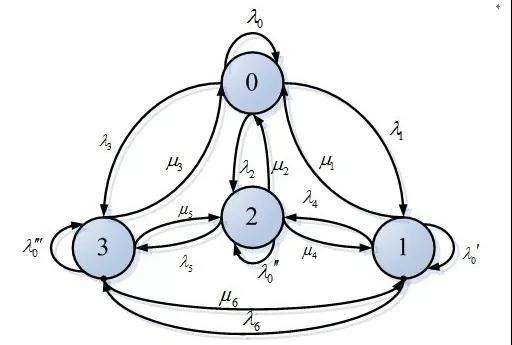


图4-6

时间和状态都是离散的马尔可夫过程称为马尔可夫链, 简记为：

**（n=0,1,2......）**

马尔可夫链是随机变量的一个数列。这些变量的范围，即他们所有可能取值的集合，被称为“状态空间”，而的值则是在时间n的状态。如果对于过去状态的条件概率分布仅是的一个函数，则



马尔可夫链的性质:









离散状态空间中的马尔可夫模型

如果状态空间是有限的，则转移概率分布可以表示为一个具有(i,j)元素的矩阵，称之为“转移矩阵”：



对于一个离散状态空间，k步转移概率的积分即为求和，可以对转移矩阵求k次幂来求得。就是说，如果是一步转移矩阵，就是k步转移后的转移矩阵。

平稳分布是一个满足以下方程的向量：



在此情况下，稳态分布 是一个对应于特征根为1的、该转移矩阵的特征向量。如果转移矩阵不可约，并且是非周期的，则收敛到一个每一列都是不同的平稳分布，并且



独立于初始分布。这是由Perron-Frobenius theorem所指出的。正的转移矩阵（即矩阵的每一个元素都是正的）是不可约和非周期的。矩阵被称为是一个随机矩阵，当且仅当这是某个马尔可夫链中转移概率的矩阵。

3.3.4 bp神经网络模型

神经网络是由大量的、简单的处理单元（称为神经元）广泛地互相连接而形成的复杂网络系统，它反映了人脑功能的许多基本特征，是一个高度复杂的非线性动力学习系统。神经网络具有大规模并行、分布式存储和处理、自组织、自适应和自学能力，特别适合处理需要同时考虑许多因素和条件的、不精确和模糊的信息处理问题。BP网络是一种对非线性可微分函数进行权值训练的多层网络。它的最大特点是仅仅借助样本数据,无需建立系统的数学模型,就可对系统实现由m个输入神经元的模式向量p组成的pm空间到空间n(为输出节点数)的高度非线性映射。

BP神经网络的计算过程由正向计算过程和反向计算过程组成。正向传播过程，输入模式从输入层经隐单元层逐层处理，并转向输出层，每一层神经元的状态只影响下一层神经元的状态。如果在输出层不能得到期望的输出，则转入反向

传播，将误差信号沿原来的连接通路返回，通过修改各神经元的权值，使得误差信号最小。

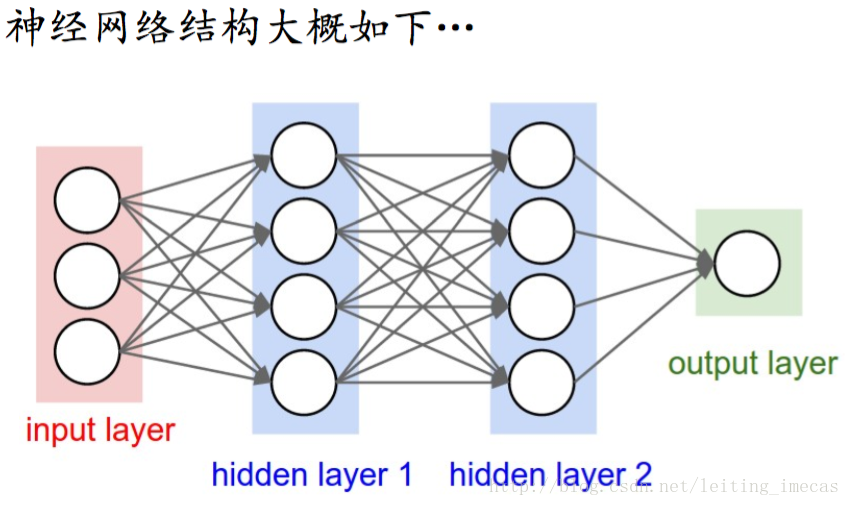


图4-7

LinearRegression模型：



sigmoid函数：



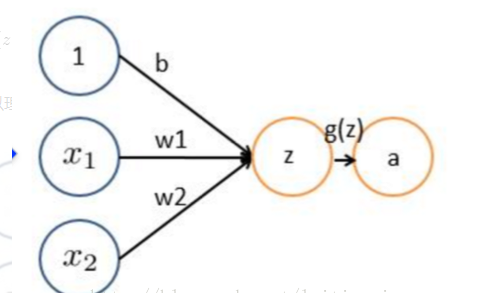


图4-8

3.3.5支持向量机模型

支持向量机是一种二分类模型，它的目的是寻找一个超平面来对样本进行分割，分割的原则是间隔最大化，最终转化为一个凸二次规划问题来求解。由简至繁的模型包括：

当训练样本线性可分时，通过硬间隔最大化，学习一个线性可分支持向量机；

当训练样本近似线性可分时，通过软间隔最大化，学习一个线性支持向量机；

当训练样本线性不可分时，通过核技巧和软间隔最大化，学习一个非线性支持向量机。

分离超平面为，如果所有的样本不光可以被超平面分开，还和超平面保持一定的函数距离（下图函数距离为1），那么这样的分类超平面是比感知机的分类超平面优的。可以证明，这样的超平面只有一个。和超平面平行的保持一定的函数距离的这两个超平面对应的向量，我们定义为支持向量。

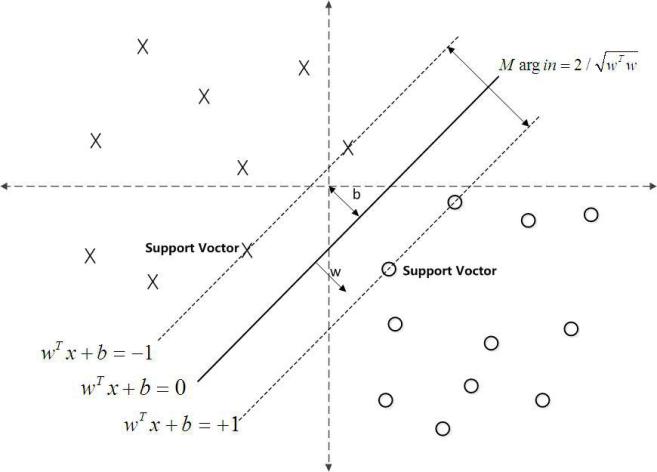


图4-9

支持向量机的目标函数：



支持向量机基于几何间隔最大化原理，认为存在最大几何间隔的分类面为最优分类面。

（3）可视化技术

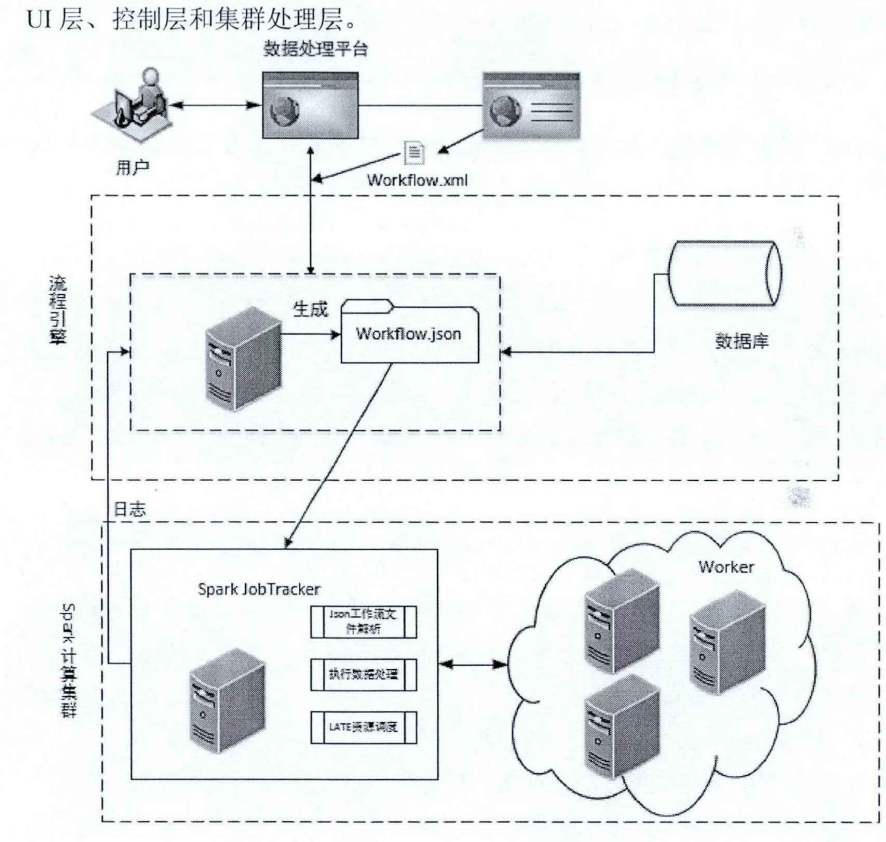
一、Spark的大数据处理流程的可视化方案

这里的第一项可视化成果是基于Spark的大数据处理可视化工具。设计者分析了当今Spark大数据处理的现状，就其问题使用可视化方法进行了解决。设计者认为当前Spark大数据处理存在两个问题：一是Spark数据处理的中间过程和结果对于普通用户而言不友好，如果用户没有接触过Spark编程，他们可能会存在阅读和操作数据上的不便。二是Spark的代码难以在不同的工程上复用，增加了数据处理的时间成本和劳动成本。设计者根据这些问题，提出了一套可视化操作系统方案，用于优化Spark数据处理的流程，提高用户与系统的交互便利。

大致来说，该方案包括了两个部分，一个是根据Spark处理流程，在Spark的控制层和展示层进行可视化设计，以达到用户与系统进行交互以及查看数据的目的。另一个则是使用可视化工具管理系统日志，以达到监控和管理系统运行的目的。

Spark的大数据处理架构展示如下图。根据架构图可将Spark数据处理系统分为三个层次：Web层、控制层、Spark数据处理层。Web层是基于各种标记语言实现的系统展示平台，用户通过该平台与系统进行交互。控制层要接受Web上用户的请求，然后由Spark集群执行获取相关结果返回到Web层，且该层实际上具备的是数据处理的任务控制流程。Spark数据处理层为控制层提供计算服务和任务调度的服务，同时也即大数据的分布式计算的主体。开发者亦是根据该体系结构来完成可视化设计的。

根据数据分析的交互要求，开发者设计了一个处理数据处理流程文件的工具。该工具基于纯Web技术设计实现，布局采用基于浏览器的div+css控制的页面技术实现，通过div元素来模拟各个布局模块，通过css叠加样式来完成页面布局。该工具共有一下几个模块构成：1.工具栏 2.编辑区 3.流程视图区 4.属性编辑区，而实现图如下所示4-10。



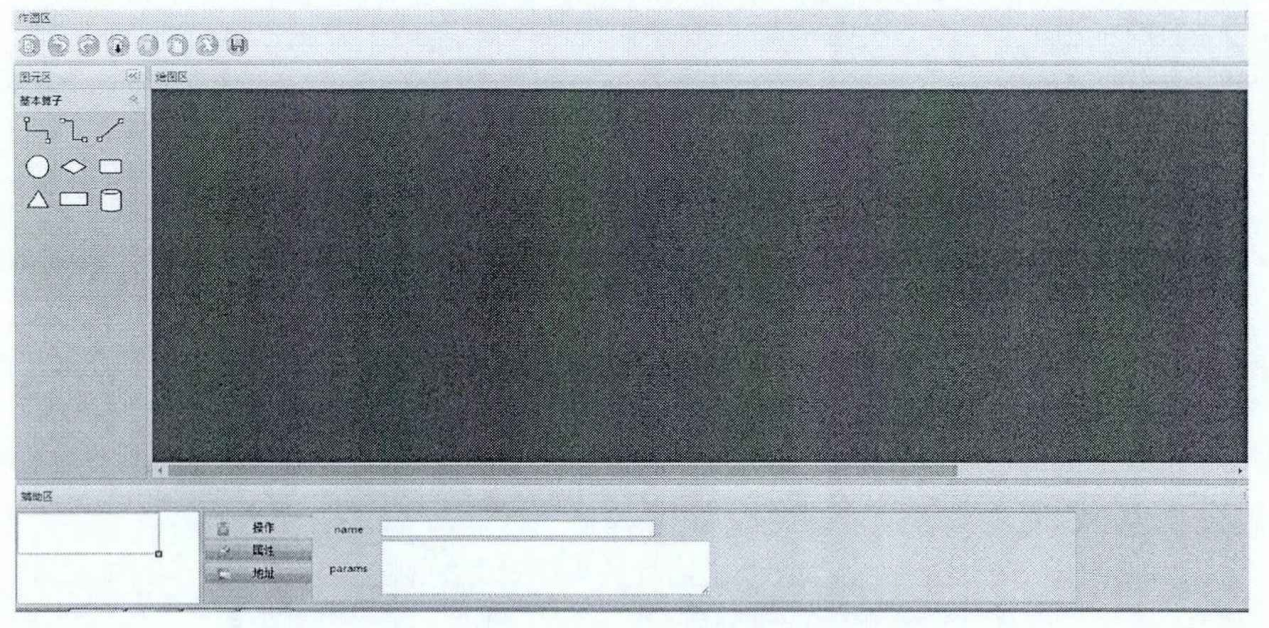


图4-10

二、图数据处理系统

大数据可视化，不同于传统的信息可视化，面临最大的一个挑战就是规模，如何提出新的可视化方法能够帮助人们分析大规模、高维度、多来源、动态演化的信息，并辅助作出实时的决策，成为了这个领域最大的挑战。不难想象股市所含有的巨量数据也包含有上述特征。为了解决这个问题，我们可以依赖的主要手段是两种，即数据转换和视觉转换。

现有研究工作主要聚焦在 4 个方面：

(1) 通过对信息流进行压缩或者删除数据中的冗余信息对数据进行简化。

(2) 通过设计多尺度、多层次的方法实现信息在不同的解析度上的展示，从而使用户可自主控制展示解析度。

(3) 利用创新的方法把数据存储在外存，并让用户可以通过交互手段方便地获取相关数据，这类研究也成为核外算法。

(4) 提出新的视觉隐喻方法以全新的方式展示数据。

第一种方案依赖基本的数据转换方法来简化数据，涉及到计算机图形学方面的知识，限于篇幅仅简要介绍。其主要做法有通过删除节点以及包含这个节点的

三角形简化数据生成的网格的，或者通过渐进网格表达方法简化图形。

第二种方法主要聚焦在地理数据的处理上，与股市数据不是有很大关联，在此简单介绍。例如，一些使用固定网格方法的系统建立在直角三角形的层次结构之上，而一些不规则三角形网则是通过不把三角形限制在固定网格上的方式来解决这个问题。也有将网格进一步扩展到四面体上的做法。

第三种方法考虑到了内存承受数据量的局限性，提出了依靠策略在外存存储数据的做法。目前人们设计了一些新的算法与分析工具来解决几何算法以及可视化方法。

第四种方法着重过滤不重要信息。一类典型的方法是“焦点+上下文”方法,它重点对焦点数据进行细节展示，对不重要数据的则简化表示，例如鱼眼视图。也有使用空间树，一种树形浏览器通过动态调整树枝的尺寸来使其最好地适配显示区域。

图数据可视化的必要手段是对图结构和交换关系网络的可视化技术，基于所谓的力驱动模型对图像进行布局。即考虑图中任意两点间都存在排斥力的情况下，去计算图中所有节点在空间中的稳定位置坐标让系统势能最小。Tutte和Eades分别在1963年和1984年，奠基了力驱动模型的布局方法。图可视化方法还与具体的数据挖掘算法有关，可视化方法和工具均基于算法的特征进行设计。目前典型的工具和方法有慕尼黑大学的PBC交互式决策树分类器、Prefuse Xmdvtool等。Huang等人在2000年提出交互式聚类算法和分类算法。

以上是图数据处理方法的主要内容，考虑到股票数据的可见性，尤其是需要展现股市的走势和各种历史变化，适当采用图形策略进行处理还是值得考虑的。

三、股市交互式可视化分析系统

根据上文的图数据处理方法，我们知道了使用图形相关的工具来处理数据可视化会得到不错的效果。一下介绍另一个基于图数据处理方法的可视化方案，该方案提出了一个支持快速选择、自动分组及导航浏览三种股市图交互式可视化操作的大规模股市图分析系统。其使用了极大团的理论，并进行了一定的优化。

该方案基于开源软件GraphStream开发其股市图交互式可视化分析系统。采用真实的数据源S&P 500数据集，基于S&P 500数据集构建的股市图可视化后的效果图如下图4-11。该系统拥有三个主要功能：快速选择、自动分组、导航浏览。快速选择能够根据用户感兴趣的一组股市图结点来快速枚举，获得相关的极大团。图4-12自动分组能够根据用户感兴趣的一组股市图节点，查看它们间的组合关系并从中得到与特定股票相关的极大团。导航浏览能够根据用户感兴趣的股市图节点间的组合关系，进一步显示相关极大团中其他的股市图节点。图4-13

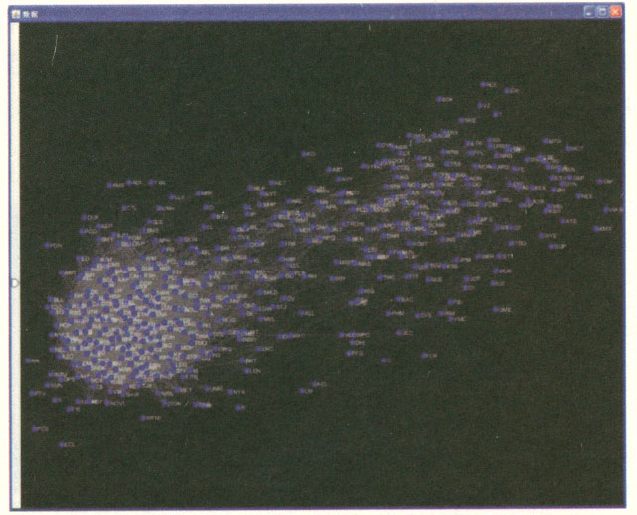


图4-11

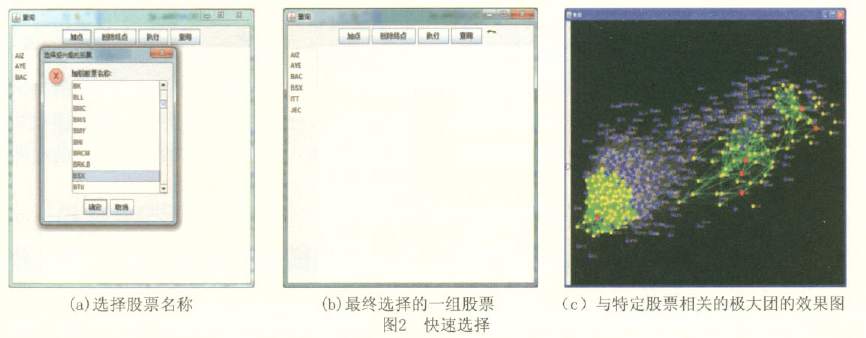


图4-12

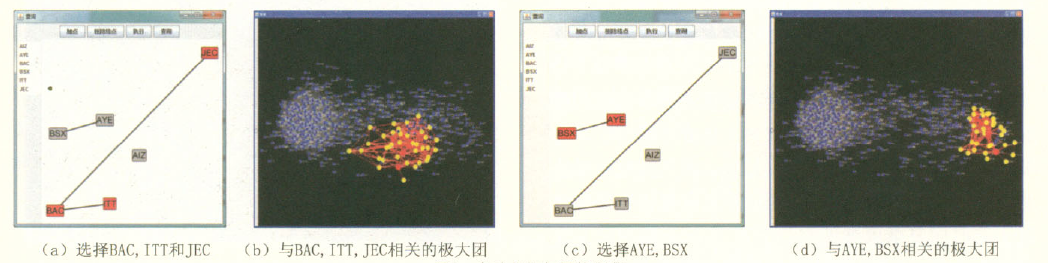


图4-13

总得来说，这是一个根据图形理论来进行数据挖掘的系统，而且它与可视化技术紧密联系在了一起，做到了将复杂数据简单地呈现给用户的目的，具有很强的应用性。此处的可视化采用了上文所描述的力驱动模型进行布局，同时其极大团枚举算法也是图理论和图算法的研究重心，随着社交网路的兴起，该领域吸引了越来越多人的主意，成为了研究的重点。

根据上文我们所得到的调研结果，在这进行其横向对比：

|  |  |  |  |
| --- | --- | --- | --- |
| 章节 | Spark的大数据处理流程的可视化方案 | 图数据处理系统 | 股市交互式可视化分析系统 |
| 数据处理核心 | Spark集群 | 主要是理论论述，没有涉及 | 基于理论自行设计的计算核心 |
| 计算理论 | 未涉及 | 未涉及 | 极大团理论 |
| 可视化方案 | 使用多种前端技术，根据分层思想收集底层数据，并呈现在Web层。同时实现用户与系统的交互 | 使用图数据处理系统，尽可能解决呈现数据的问题 | 设计一个交互式可视化系统，降低用户使用数据的门槛 |
| 特点 | 重视分层次设计，倾向于使用软件设计模式解决可视化难题 | 重视数据图像化，使用计算机视觉技术突破可视化难题 | 将数据图像化与交互环节相结合，提高可读性和操作性 |
| 优点 | 具备成熟的软件理论，可视化方案拓展性强 | 数据可读性强 | 数据可读性强，且用户使用门槛不高 |
| 缺点 | 程序冗余程度高，没有将性能作为第一要素考虑 | 缺乏拓展性，且内核缺乏巨量数据处理能力 | 缺乏拓展性 |

**五、可使用技术补充**

1. Docker容器技术

Docker 是一个开源的应用容器引擎，让开发者可以打包他们的应用以及依赖包到一个可移植的容器中，然后发布到任何流行的 Linux 机器上，也可以实现虚拟化。容器是完全使用沙箱机制，相互之间不会有任何接口。

作为一种轻量级的虚拟化方式，Docker在运行应用上跟传统的虚拟机方式相比具有显著优势：

Docker容器很快，启动和停止可以在秒级实现，这相比传统的虚拟机方式要快得多。

Docker容器对系统资源需求很少，一台主机上可以同时运行数千个Docker容器。

Docker通过类似Git的操作来方便用户获取、分发和更新应用镜像，指令简明，学习成本较低。

Docker通过Dockerfile配置文件来支持灵活的自动化创建和部署机制，提高工作效率。



图5-1

（2）TensorFlow

TensorFlow 是一个开放源代码软件库，用于进行高性能数值计算。借助其灵活的架构，用户可以轻松地将计算工作部署到多种平台（CPU、GPU、TPU）和设备（桌面设备、服务器集群、移动设备、边缘设备等）。TensorFlow 最初是由 Google Brain 团队（隶属于 Google 的 AI 部门）中的研究人员和工程师开发的，可为机器学习和深度学习提供强力支持，并且其灵活的数值计算核心广泛应用于许多其他科学领域.

TensorFlow的主要特征：

高度的灵活性。只要可以将计算表示为一个数据流图，就可以使用Tensorflow来构建图，描写驱动计算的内部循环。

真正的可移植性。Tensorflow 在CPU和GPU上运行。

自动求微分。基于梯度的机器学习算法会受益于Tensorflow自动求微分的能力。只需要定义预测模型的结构，将这个结构和目标函数（objective function）结合在一起，并添加数据，Tensorflow将自动计算相关的微分导数。

多语言支持。虽然TensorFlow拥有合理易用的C++和Python使用界面，但也支持其他语言的使用，如Go，Java，Lua，Javascript或R。

性能最优化。TensorFlow 给予了线程、队列、异步操作等以最佳的支持，TensorFlow 让硬件的计算潜能全部发挥出来。

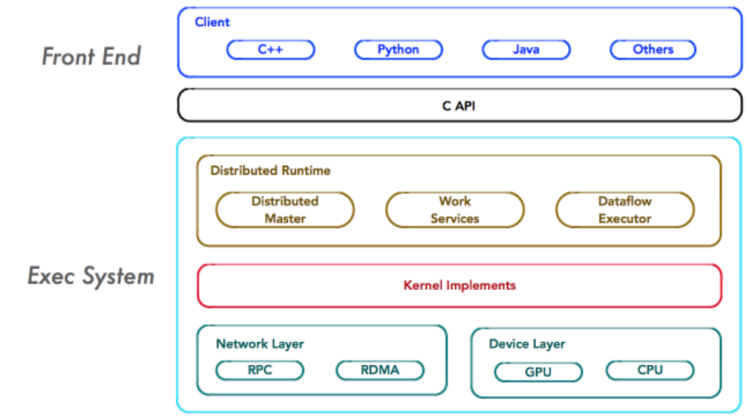


图5-2

**六、实施计划**

本项目可根据软件开发的流程划分为以下几个阶段，即：项目开题、需求分析、概要设计、详细设计、程序版本1.0实现、程序版本2.0实现、集成测试、验收测试及系统维护等9个阶段。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **阶段** | **主要工作** | **阶段可交付物** | **计划起始时间** | **计划终止时间** |
| 项目开题  (2018/10/19-  2018/11/9) | 调研用户需求及用户环境 |  | 2018/10/19 | 2018/11/3 |
| 论证项目可行性 |  | 2018/10/30 | 2018/11/3 |
| 制定项目初步计划 | 项目初步开发计划进度表 | 2018/11/3 | 2018/11/9 |
| 编写开题报告 | 开题报告 | 2018/10/30 | 2018/11/9 |
| 需求分析  (2018/10/19-  2018/11/15) | 确定集群运行环境 | 需求规格说明 | 2018/10/19 | 2018/11/3 |
| 建立系统逻辑模型 | 项目开发计划 | 2018/11/3 | 2018/11/15 |
| 确定系统功能及性能要求 |  | 2018/10/19 | 2018/11/9 |
| 编写需求规格说明、用户手册概要、测试计划 | 用户手册概要 | 2018/11/3 | 2018/11/9 |
| 确认项目开发计划 |  | 2018/11/9 | 2018/11/15 |
| 概要设计  (2018/11/15-  2018/11/30) | 建立系统总体结构，划分功能模块 |  | 2018/11/15 | 2018/11/22 |
| 定义各功能模块接口 |  | 2018/11/15 | 2018/11/22 |
| 数据仓库设计 | 数据仓库设计说明书 | 2018/11/22 | 2018/11/25 |
| 制定组装测试计划 |  | 2018/11/25 | 2018/11/27 |
| 对已完成的文档进行评审 |  | 2018/11/29 | 2018/11/30 |
| 详细设计  (2018/11/25-  2018/12/20) | 设计各模块实现的算法与模型 |  | 2018/11/25 | 2018/12/02 |
| 确定模块间详细接口 |  | 2018/12/03 | 2018/12/15 |
| 制定模块测试方案 |  | 2018/12/16 | 2018/12/20 |
| 程序版本1.0开发  (2018/12/21-  2019/2/28) | 编写程序版本1.0源代码 |  | 2018/12/21 | 2019/01/31 |
| 进行模块测试和调试 | 用户手册 | 2019/2/1 | 2019/2/15 |
| 编写用户手册 |  | 2019/2/15 | 2019/3/20 |
| 形成中期检查报告 |  | 2019/2/20 | 2019/2/28 |
| 程序版本2.0开发（2019/3/1-2019/3/25） | 编写程序版本2.0源代码 |  | 2019/3/1 | 2019/3/15 |
| 进行模块测试和调试 | 用户手册 | 2019/3/16 | 2019/3/25 |
| 集成测试  (2019/3/26-  2019/4/4) | 执行集成测试计划 |  | 2019/3/24 | 2019/3/27 |
| 编写集成测试报告 |  | 2019/3/27 | 2019/4/4 |
| 验收测试  (2019/4/5-  2019/4/15) | 测试整个软件系统（健壮性测试） |  | 2019/4/5 | 2019/4/10 |
| 试用用户手册 | 用户手册 | 2019/4/10 | 2019/4/13 |
| 编写开发总结报告 | 开发工作总结 | 2019/4/10 | 2019/4/15 |
| 维护  (2019/4/15-  2019/5/1) | 为纠正错误，完善应用而进行修改 |  | 2019/4/15 | 2019/4/30 |
| 对修改进行配置管理 |  | 2019/4/15 | 2019/4/30 |
| 编写故障报告和修改报告 |  | 2019/4/25 | 2019/4/30 |
| 修订用户手册 |  | 2019/4/25 | 2019/4/30 |

参卡文献：

[1] Kalev P S., Liu W M., Pham P K., et al. Public information arrival and volatility of intraday stock returns[J]. Journal of Banking & Finance, 2004,28(6):1441-1467

[2] Mittermayer MA. Forecasting intraday stock price trends with text mining techniques[C].System Sciences, 2004. Proceedings of the 37th Annual Hawaii InternationalConference on IEEE,2004:10 pp

[3] Ryan P., Taffler R J. Are Economically Significant Stock Returns and Trading Volumes Driven by Firm specific News Releases[J]. Journal of Business Finance & Accounting,2004, 31(1-2): 49-82

[4]Wuthrich B., Cho v., Leung S., et al. Daily stock market forecast from textual web data[C] Systems, Man, and Cybernetics, 1998. 1998 IEEE International Conference on IEEE,1998, 3: 2720-2725

[5]Li F. The Information Content of Forward Looking Statements in Corporate Filings-A Naive Bayesian Machine Learning Approach[J]. Journal of Accounting Research, 2010, 48(5):1049-1102

[6]Antweiler w., Frank M Z. Is all that talk just noise? The information content of internet

[7]Tetlock P c. Giving content to investor sentiment: The role of media in the stock market[J]. The Journal of Finance, 2007, 62(3): 1139-1168