

硕士研究生学位论文

|  |  |
| --- | --- |
| 题目： | 面向微博的在线广告点击 |
|  | 行为预测系统设计与实现 |

|  |  |
| --- | --- |
| 姓 名： | 孙超 |
| 学 号： | 1301221193 |
| 院 系： | 软件与微电子学院 |
| 专 业： | 软件工程 |
| 研究方向： | 互联网信息挖掘与处理 |
| 导师姓名： | 王海峰、刘宏志 |

二〇一 年 月

版权声明

任何收存和保管本论文各种版本的单位和个人，未经本论文作者同意，不得将本论文转借他人，亦不得随意复制、抄录、拍照或以任何方式传播。否则，引起有碍作者著作权之问题，将可能承担法律责任。

# 摘要

广告点击行为预测是互联网计算广告的重要环节，通常的做法是先用机器学习算法在已有用户的特征数据和点击行为记录上训练出一个模型，然后用训练出的模型结合新的用户特征数据返回该用户点击行为的概率。以往工作的重点是训练模型参数选取和用户特征选取上，其中用户特征选取多集中于结构化特征，即如用户的性别、年龄、职业、教育程度等固有属性。

本文在用户结构化特征基础上，利用用户的微博行为数据（包括发表、转发、点赞）过的文章，通过特征工程转换为用户的非结构化特征，用GBDT和LR融合的方法做点击行为预估，并设计和实现了一个可商业化运营的线上分布式系统。实验和线上运行记录表明该方法较传统的只用结构化特征的方法在预估准确率方面有了明显的提高。

关键词：广告，点击率，预测

# ABSTRACT

Click-through rate estimation plays a key role in computational advertising. The usual approach is to use machine learning algorithm to train a model based on users' featured data, and then use this model to predict a new user's click behavior. In the past, most of works focus on selecting the right arguments and features of users.（待补充）

KEY WORDS: Key word 1, Key word 2, Key word 3, ……

# 目录

[摘要 I](#_Toc475755295)

[ABSTRACT II](#_Toc475755296)

[目录 III](#_Toc475755297)

[第一章 绪论 1](#_Toc475755298)

[1.1 项目背景 1](#_Toc475755299)

[1.2 广告点击率预测研究发展现状 1](#_Toc475755300)

[1.3 本文的研究内容和方法 2](#_Toc475755301)

[1.4 本文内容组织结构 3](#_Toc475755302)

[第二章 相关技术 4](#_Toc475755303)

[2.1 逻辑回归模型 4](#_Toc475755304)

[2.1.1 逻辑函数 5](#_Toc475755305)

[2.1.2 模型调整 6](#_Toc475755306)

[2.2 在线学习算法FTRL 7](#_Toc475755307)

[2.2.1 背景 7](#_Toc475755308)

[2.2.2 相关算法 8](#_Toc475755309)

[2.2.3 FTRL算法实现 9](#_Toc475755310)

[2.2.4 算法优化 10](#_Toc475755311)

[2.3 迭代决策树 GBDT 10](#_Toc475755312)

[2.3.1 回归决策树 Regression Decision Tree 11](#_Toc475755313)

[2.3.2 梯度迭代（Gradient Boosting） 11](#_Toc475755314)

[2.3.3 GBDT工作过程 12](#_Toc475755315)

[2.3.4 Shrinkage 12](#_Toc475755316)

[2.4 GBDT和LR结合用于CTR预测 13](#_Toc475755317)

[2.5 在线广告点击预测系统构成 14](#_Toc475755318)

[2.5.1 HTTP服务器Ngnix 14](#_Toc475755319)

[2.5.2 分布式配置和集群管理工具ZooKeeper 14](#_Toc475755320)

[2.5.3 全文检索引擎Lucene 15](#_Toc475755321)

[2.5.4 跨语言通信接口Thrift 15](#_Toc475755322)

[2.5.5 数据高速公路Flume 16](#_Toc475755323)

[2.5.6 分布式数据处理平台Hadoop 16](#_Toc475755324)

[2.5.7 特征在线缓存Redis 17](#_Toc475755325)

[2.6 本章小结 17](#_Toc475755326)

[第三章 系统需求分析 18](#_Toc475755327)

[3.1 性能需求 18](#_Toc475755328)

[3.2 功能需求 19](#_Toc475755329)

[3.3 非功能需求 19](#_Toc475755330)

[第四章 系统设计与实现 20](#_Toc475755331)

[4.1 系统整体架构 20](#_Toc475755332)

[4.2 Service定义 20](#_Toc475755333)

[4.3 ApiServer的设计与实现 22](#_Toc475755334)

[4.3.1 Service管理模块 22](#_Toc475755335)

[4.3.2 请求处理模块 23](#_Toc475755336)

[4.3.3 算法服务管理模块 24](#_Toc475755337)

[4.3.4 负载均衡和异常处理 26](#_Toc475755338)

[4.4 算法服务的设计与实现 26](#_Toc475755339)

[4.5 本章小结 27](#_Toc475755340)

[第五章 业务流程 28](#_Toc475755341)

[5.1 微博文章处理 28](#_Toc475755342)

[5.1.1 Jieba分词 28](#_Toc475755343)

[5.1.2 文章向量化 28](#_Toc475755344)

[5.2 GBDT特征工程 29](#_Toc475755345)

[5.3 LR预估 30](#_Toc475755346)

[5.4 本章小结 30](#_Toc475755347)

[第六章 系统测试 31](#_Toc475755348)

[6.1 线下测试 31](#_Toc475755349)

[6.2 线上环境测试 32](#_Toc475755350)

[第七章 总结与展望 33](#_Toc475755351)

[7.1 总结 33](#_Toc475755352)

[7.2 展望 33](#_Toc475755353)

[参考文献 34](#_Toc475755354)

[致谢 36](#_Toc475755355)

[北京大学学位论文原创性声明和使用授权说明 37](#_Toc475755356)

# 第一章 绪论

## 1.1 项目背景

当前我国互联网产业迅速发展，根据中国互联网络信息中心（CNNIC）的最新统计，截至2016年6月，我国网民规模达7.10亿，互联网普及率达到51.7%，与2015年底相比提高1.3个百分点，超过全球平均水平3.1个百分点，超过亚洲平均水平8.1个百分点[1]。与此同时，电子商务产业也发展迅速，十二五期间，中国的网络零售交易额规模跃居世界第一，网购在网民中的普及率高达55.7%，仅在2016年双11期间，天猫交易额就突破1200亿元，物流订单量超6.5亿[2]。

在此背景下催生了互联网广告这一新兴产业，互联网广告经历了从最初铺天盖地的横幅广告到人群及兴趣精准定向的搜索广告与推荐引擎，进而催生了计算广告学这一新兴研究课题，这一课题涉及大规模搜索和文本分析、信息获取、统计模型、机器学习、分类、优化以及微观经济学等诸多领域的知识[3]。相比于基于传统媒体如电视、广播、报纸海量而无特定目标的广告投放，以互联网为媒介的广告投放可以根据用户特征（主要包括用户个人信息和上网行为）做到个性化精准投放。从而为大大提高了广告收益，为企业节省了推广成本。其中，广告点击率预测(Click Through Rate 简称CTR)是计算广告学的一个重要研究内容，是互联网企业广告竞价排名和流量变现业务的重要支撑。

## 1.2 广告点击率预测研究发展现状

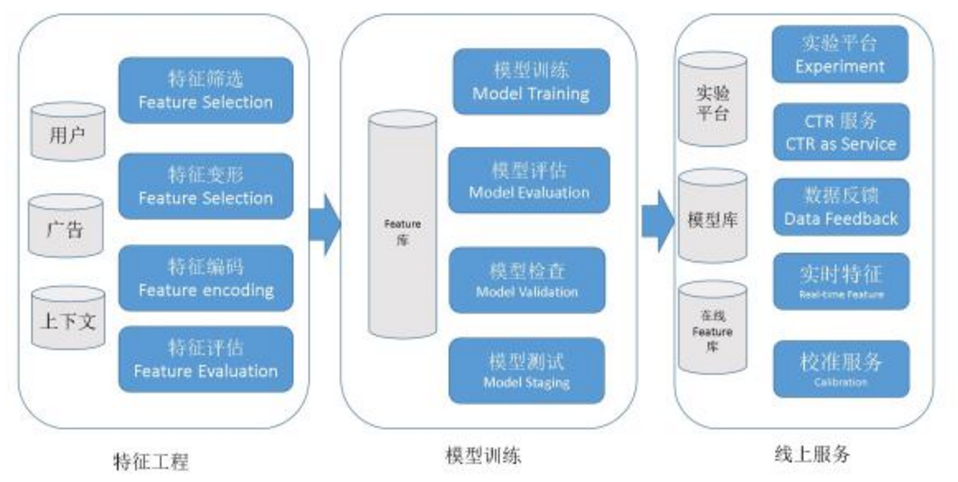
目前广告点击行为预测研究主要集中于特征学习、数据特点和用户行为这3个方面，本文将以这3个方面分别介绍已有的相关工作。

特征学习方面，影响CTR预估准确率的特征很多，不是选用的特征越多准确率就会越高，如何科学的选取又用的特征是不少学者和工程师的研究方向。在这方面，文献[4]研究了原始查询和重写后查询之间的相关性与广告点击率之间的关系，该文章考察了一些特征如次序、长度差异、编辑距离等对广告点击率的影响；文献[5][6]将广告显示位置和广告查询相关性作为特征，同时根据相似广告来解决非常见广告和新广告的点击率预测问题；文献[7]综合运用协同过滤、贝叶斯网络和特征工程等模型来预测点击率；文献[8]提出用协同过滤和张量分解来提取用户个性化特征用于点击率预测；

数据特点方面，广告数据多为稀疏数据，文献[4]提出利用相似广告来预估目标广告的点击率；针对新广告的历史数据缺失问题，文献[9]提出“竞拍词-广告主”矩阵，用层次聚类的方法解决历史数据不充分的问题；文献[10]设计了用于稀疏广告和新广告的点击率预测模型，提出了基于层级结构的预估模型和基于Time-Spatial的预估模型；文献[11]提出了基于经验贝叶斯的自然数据分层和基于数据一致性的两种平滑计算方法对层级模型做了改进。

用户行为建模，用假设检验和贝叶斯网络刻画用户浏览场景，进而估计出用户点击广告的概率。文献[12]基于用户的浏览行为假设提出了级联模型，如果用户点击一个文档，若该文档不能满足用户的查询需求，则该用户很可能继续查看后续搜索结果并有点击行为； 文献[13]扩展级联模型到多次点击。

## 1.3 本文的研究内容和方法

广告点击率预测过程可分为3个步骤：特征工程、魔性训练和线上服务。

在这其中，特征工程是很重要的一个环节，特征选取的优劣对最终预测结果的准确率有很大的影响。对于用户而言，用户的固有信息如年龄、性别、地域、职业、教育程度等信息可作为用户特征，我们称之为结构化特征；而如用户的行为数据如浏览过的网页、购买过的物品、转发过的微博等非固有的，随时间和环境变化的特征，我们称之为非结构化特征。而什么样的特征适合做CTR预估，业界并没有统一的标准。往往靠工程科研人员结合具体情况和生活常识通过反复实验和以往经验来决定。以年龄特征为例：年轻人普遍喜欢运动类的广告，30岁左右的男人喜欢车，房子之类的广告，50岁以上的人喜欢保健品的广告。再以性别特征为例：化妆品的广告在女性上面的点击率就比在男性上面的点击率高很多，又如体育用品的广告在男性上面的点击率也比女性高，说明性别这个特征在化妆品和体育行业也是有预测能力的，经过多个行业的验证，就认为性别这个特征可以用了。在实际的使用中发现，性别这个特征比较有效，手机平台这个特征也比较有效，地域和年龄这两个特征有一定效果，但没有前两个那么明显，跟它们的使用方式可能有关。以上示例说明在CTR预估上特征选择往往根据具体应用场景的，主观的。

CTR预估中用的最多的是逻辑回归模型，又称LR(Logistic Regression)模型，是一种广义线性模型。LR使用了Logit变换将函数值映射到0~1区间，映射后的函数值就是CTR的预估值。LR模型的优点是效率高，容易并行化，处理上亿级别数据不是问题。但缺点是作为一种线性模型，其学习能力有限，需要大量特征工程预先分析出有效的特征、特征组合，从而去间接增强LR 的非线性学习能力[14]。上文中提到，特征选取往往依靠人工经验，耗时耗力且不一定能带来效果提升。如何自动发现有效的特征、特征组合，弥补人工经验不足，Facebook在2014年的文章介绍了通过GBDT(Gradient Boost Decision Tree)解决LR的特征组合问题[15]，随后Kaggle竞赛将此付诸实践[16]。在此之后GBDT与LR融合的方法开始引起了业界关注。

本文在结构化特征的基础之上，利用用户的微博行为记录(包括发表、转发、点赞等行为)，使用GBDT方法在用户有过行为的微博文章里提取非结构化特征，与结构化特征一起作为LR模型的训练数据训练出线上模型。实验证明，用了此方法后CTR预估准确率较之前只使用结构化特征有了明显的提升。此外，本文还详细介绍了基于此种方法的分布式在线实时广告点击率预测系统的设计和实现。

## 1.4 本文内容组织结构

第一章：绪论，介绍了广告点击率预测背景、发展现状和本文的主要工作，最后总结文章的组织结构。

第二章：相关技术，

# 第二章 相关技术

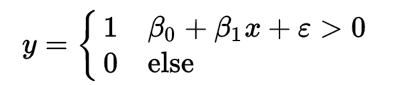
## 2.1 逻辑回归模型

逻辑回归(Logistic Regression)模型，是机器学习中的一种分类模型，由于算法的简单和高效，在社会学、生物统计学、临床、数量心理学、计量经济学、市场营销等领域得到广泛的应用。

逻辑回归算法最早由统计学家David Cox提出[17][18]，逻辑回归有二分类、多分类和顺序分类[19] 。二分类模型是应用最多的模型，其最终输出结果值只有两种，如“生”或“死”，“胜利”或“失败”，在本文的应用场景里就是用户“点击”或“未点击”广告，通常这两种结果值用“0”和“1”来表示[20] ，本文中“0”表示未点击，“1”表示点击。逻辑回归被用作预测基于相互独立变量基础上某一事件发生的几率，

和其他的回归分析方法类似，逻辑回归使用一个或多个预测变量，这些变量可以是连续的也可以是离散的。和一般线性回归不同的是，逻辑回归是用来预测相互依赖的二项变量而非连续值结果。而这违反了线性回归的假设前提，尤其是残差不服从正态分布。此外，线性回归可能对相互依赖的二项分布的变量做出无意义的预测。因此需要一种方法将二项分布值转换为连续实数值。要做逻辑回归首先用相互独立的变量表示不同等级事件发生的几率，然后计算出这些几率所占比率，然后再对这些计算出来的几率取对数，这一步称作对数变换，被看作是逻辑回归的连接函数。

对数函数的结果被用作线性回归分析预测，预测的结果通过自然对数的反函数（指数函数）转换为事件发生几率。因此，尽管观察到的逻辑回归预测结果是0-1值，但逻辑回归实在预测一个连续值得概率。在有些应用场合中，这个连续的概率值就是最终所需；在其他应用场景则需要一个或是或非的布尔值，通过定义一个阈值将连续值转换为离散的布尔值。

逻辑回归可被看作是寻找一组最有的β参数以满足：  
这里ε是逻辑斯谛分布误差。这里定义了一个隐变量 ，误差项ε未被测量，因此变量y’也是不可估的，因此被称作隐变量。与普通回归分析不同的是，β参数不可以直接用y和x的方程式表示，相反，需要用反复迭代过程寻找这样一个表达式，这正是逻辑回归训练算法要做的工作。

### 2.1.1 逻辑函数

逻辑函数是逻辑回归模型的重要组成部分，逻辑函数之所以有用是因为它的输入可以是任意实数，而输出总是在(0,1)区间上，因此比较适合用来表示概率。逻辑函数定义如下：  
函数图形如下图所示：  
假设t是x的线性函数，t可以用x表示为：  
这样，逻辑函数就可以写成：  
F(x)可被看作最终预测的概率值。  
逻辑函数的反函数定义为：

两边取乘方得到：

其中g是对数函数，g(F(x))表明这个对数函数等价于线性回归表达式。F(x)表示因变量所代表事件发生的概率等于线性回归表达式的逻辑函数值。值得注意的是线性回归式的值域是全体实数，但是经过F(x)映射之后变为(0,1)；是线性回归方程式的截距；是回归系数。

因变量所表示的概率等同于线性回归式的指数函数。这恰好证明了对数算子是概率和线性回归式的连接函数。因为对数函数的值域是全体实数，因此它提供了足够的尺度使得在之基础之上进行线性回归。因此定义因变量的几率为：

对于连续的自变量，定义几率比(odds ratio)为：

这种指数关系解释了的作用：当x每增加1几率便乘以倍。如果因变量服从二项分布，那么几率比可定义为，这里a,b,c,d是2x2列联表的元素[21]。

### 2.1.2 模型调整

回归系数的预估一般用最大似然估计，最大似然估计也称最大概率估计，是用来估计一个概率模型参数的一种方法，其基本原理是：由已知实验结果（样本）出发，寻找一组满足这组样本分布的参数，把可能性最大的那组作为真实的参数。

设为随机变量，是多维参数向量，若随机变量相互独立且都服从，则可以得到概率函数为：

在固定时，以上函数式表示的概率；当已知时，它又变成的函数，记为：

称此函数为似然函数。似然函数值的大小意味着该样本值出现的可能性大小，由于已经得到样本值，可以估计似然函数值是比较大的，即出现上述样本可能性是高的，因此最大似然估计就是选择使达到最大值得那个作为真实模型的参数的估计。要寻找使最大的，只需在上对求导，令导函数等于0，为了求导方便，对式两边取对数得到：

对求导得到关于的k个方程组成的方程组：

解此方程组得到使函数最大的一组。

与带正态分布残差的线性回归不同的是，要寻找一个系数值的闭合表达式去最大化似然函数是不可能的，因此必须用多次迭代过程，如牛顿法[22]，该方法从一个尝试解开始在每次迭代过程中做修正看是否比之前有准确度提升直到准确度达到稳定值，称之为收敛[23]。

## 2.2 在线学习算法FTRL

FTRL(Follow The Regularized Leader)算法是由Google与2011年提出的在线学习算法[24]，在处理逻辑回归问题上有出色的性能，已在互联网行业尤其是CTR预测中得到广泛的应用。

### 2.2.1 背景

机器学习中常见的loss函数+正则化的结构风险最小化的优化问题（逻辑回归也是用的这种形式）可用两种形式描述，第一种是无约束优化形式，

第二种是带约束项的凸优化：

当合理地选择g时，二者是等价的。而相关的算法都是建立在不同的描述基础之上。

首先看批量(batch)算法，批量算法中每次迭代用全体训练数据集计算全局参数，如全局梯度。优点是能够获得较好的精度，但缺点是处理大数据集时效率低，计算开销大，而且无法处理在线环境中的数据流。对于无约束优化形式，最常见的算法是全局梯度下降法：

每次迭代求一个目标函数的全局梯度，用于计算学习率。此外还有牛顿法（切线近似）、BFGS（割线拟牛顿）等方法也是基于无约束优化的；而对于带约束项的凸优化形式，有投影梯度下降法，每次迭代后迭代结果可能位于约束集合之外，取该迭代结果在约束凸集合上的投影作为新的迭代结果，如图所示：

再来看在线算法，在线学习算法的特点是：每来一个训练样本，就用该样本产生的误差和梯度对模型进行一次修正，因此可以处理大数据量训练和在线训练。常用的算法有在线梯度下降（OGD）和随机梯度下降（SGD）等，本质思想是对单个数据的做梯度下降，因为每一步的方向并不是全局最优的，所以整体呈现出来的会是一个看似随机的下降路线。典型迭代公式如下：

这里使用混合正则化项：

其中是单点的未加和的loss函数梯度，投影集合C是约束空间。

梯度下降法的优点是精度高，但缺点主要有：简单的在线梯度下降很难产生真正稀疏的解，稀疏性在机器学习中很重要，尤其在工程应用中稀疏的特征会大大减少预测时的时间和空间复杂度；另外，梯度下降发对于不可微点的迭代会存在问题。

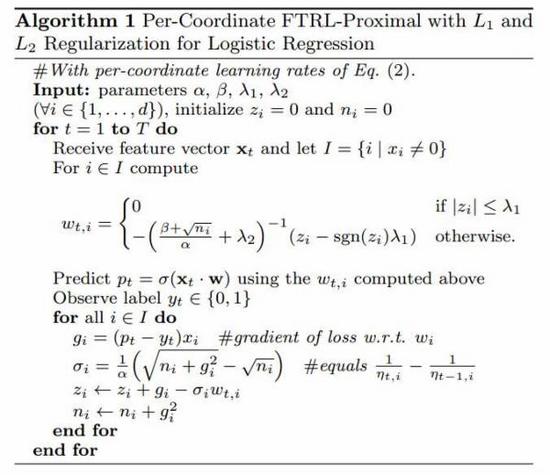
### 2.2.2 相关算法

上文提到稀疏性在机器学习中的重要性，解决此问题比较著名的方法有FOBOS(Forward-Backward Splitting method)[25]以及 RDA(Regularized dual averaging)[26]，FTRL综合了这两种算法的优点。

FOBOS是google和伯克利2009年的工作，方法与投影梯度法类似，不同之处在于将每一个数据的迭代过程分解成一个经验损失梯度下降迭代和一个最优化问题，其中最优化问题有两项：一是2范数项，不能离第一步loss损失迭代结果太远；二是正则化项，用来限定模型复杂度抑制过拟合和做稀疏化，从而保证了最终结果的稀疏性和理论上的完备。

RDA是微软2010年的工作，其主要特点是：用primal-dual algorithmic schema做非梯度下降；能够更好地在精度和稀疏性之间做平衡。

### 2.2.3 FTRL算法实现

 Google于2013年给出了逻辑回归下的per-coordinate FTRL\_Proximal算法[27]，结合了FOBOS和RDA的优点，伪代码如下：

其中per-coordinate表示 FTRL是对w每一维分开训练更新的，每一维使用的是不同的学习速率。与所有维度使用统一的学习速率相比，这种方法考虑了训练样本本身在不同特征上分布的不均匀性，如果包含w某一个维度特征的训练样本很少，那么该特征维度对应的训练速率可以独自保持比较大的值，没来一个包含该特征的样本，就可以在该样本的梯度上前进一大步，而不需要与其他特征维度的前进步调强行保持一致。

### 2.2.4 算法优化

在工程应用中会从不同的方面对算法做出一些优化来提高效率，主要优化措施有以下几种：

1). 内存节省，主要方法有：在线丢弃训练数据中很少出现的特征；对某一维度特征所来的训练样本，以一定的概率接受并更新模型；用bloom filter从概率上做某一特征出现k次才更新。

2). 浮点数重新编码，不用计算机中常用的32bit(C/C++里的float)或64bit(C/C++里的double)来存储浮点数，改用16bit编码，但是要注意处理rounding技术对regret带来的影响。

3). 训练若干相似model，对同一份训练数据序列，同时训练多个相似的model，这些model有各自独享的一些feature，也有一些共享的feature，因为有的特征维度可以是各个模型独享的，而有的各个模型共享的特征，可以用同样的数据训练。

4). Single Value Structure，多个model公用一个feature存储（例如放到cbase等），各个model都更新这个共有的feature结构，对于某一个model和它所训练的特征向量的某一维，直接计算迭代结果并与旧值做平均。

5). 训练数据重采样，在实际应用中CTR远小于50%，因此正样本更加有价值。通过对训练数据集进行重采样，可以大大减小训练数据集的大小。一般采样方法是正样本（至少有一个广告被点击）全部采样，负样本（完全没有广告被点击）按一定的比例r采样。但是直接在这种方法采样的数据集上进行训练会导致较大的预测偏差，解决方法是先采样减少负样本数目，在训练的时候再用权重弥补负样本，该权重直接乘到loss上面，从而梯度也会乘以这个权重：

## 2.3 迭代决策树 GBDT

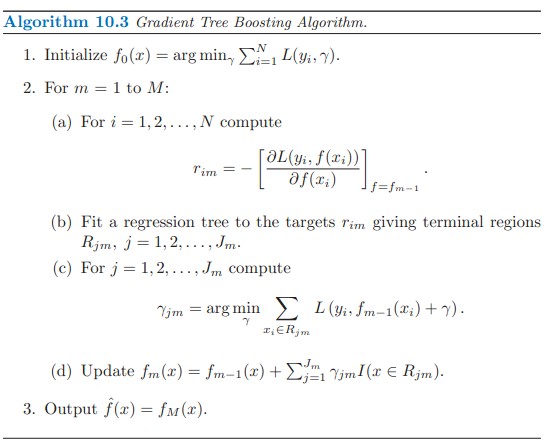
GBDT（Gradient Boost Decision Tree）是一种迭代的决策树算法，该算法由多棵决策树组成，每次迭代都在减少残差的梯度方向新建立一颗决策树，迭代多少次就会生成多少颗决策树，所有树的结论累加起来做最终答案。它在被提出之初就和SVM一起被认为是泛化能力较强的算法。近些年更因为被用于搜索排序的机器学习模型而引起大家关注。GBDT主要由三个概念组成：Regression Decistion Tree（即DT)，Gradient Boosting（即GB)，Shrinkage (算法的一个重要演进分枝，目前大部分实现都是依据该算法）。

### 2.3.1 回归决策树 Regression Decision Tree

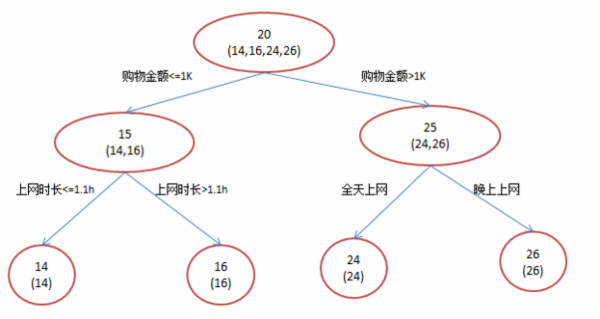
在机器学习和人工智能中，决策树可分为两大类，回归树和分类树。前者用于预测实数值，如明天的温度、用户的年龄、网页的相关程度；后者用于分类标签值，如晴天/阴天/雾/雨、用户性别、网页是否是垃圾页面。前者的结果加减是有意义的，后者则无意义。GBDT的核心在于累加所有树的结果作为最终结果，而分类树的结果是没办法累加的，所以GBDT中的树都是回归树，不是分类树。

### 2.3.2 梯度迭代（Gradient Boosting）

Boosting迭代，即通过迭代多棵树来共同决策。GBDT的核心就在于，每一棵树学的是之前所有树结论和的残差，这个残差就是一个加预测值后能得真实值的累加量。比如A的真实年龄是18岁，但第一棵树的预测年龄是12岁，差了6岁，即残差为6岁。那么在第二棵树里我们把A的年龄设为6岁去学习，如果第二棵树真的能把A分到6岁的叶子节点，那累加两棵树的结论就是A的真实年龄；如果第二棵树的结论是5岁，则A仍然存在1岁的残差，第三棵树里A的年龄就变成1岁，继续学习。

Gradient Boosting算法如下：

### 2.3.3 GBDT工作过程

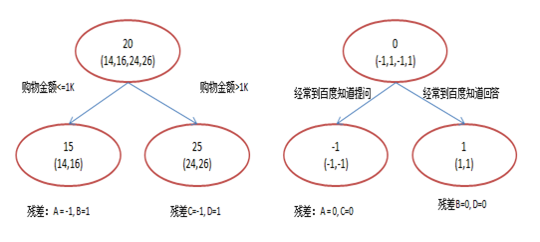
以年龄预测为例，假设训练集只有4个人A,B,C,D，他们的年龄分别是14,16,24,26。其中A,B分别是高一和高三学生；C,D分别是应届毕业生和工作两年的员工。若用一棵传统的回归决策树来训练，过程如图所示：

如果使用GBDT来来训练，过程如下：（限定叶子节点做最多有两个，并且限定只学两棵树）

A: 14岁高一学生，购物较少，经常问学长问题；预测年龄A = 15 – 1 = 14

B: 16岁高三学生；购物较少，经常被学弟问问题；预测年龄B = 15 + 1 = 16

C: 24岁应届毕业生；购物较多，经常问师兄问题；预测年龄C = 25 – 1 = 24

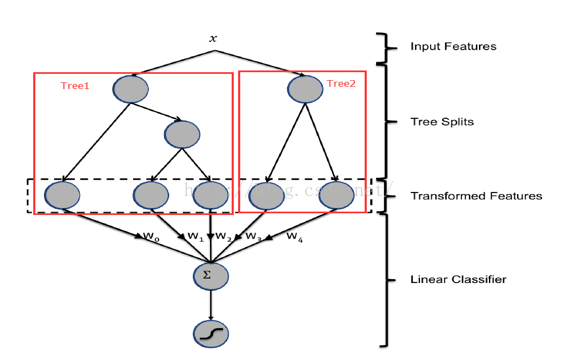
D: 26岁工作两年员工；购物较多，经常被师弟问问题；预测年龄D = 25 + 1 = 26

### 2.3.4 Shrinkage

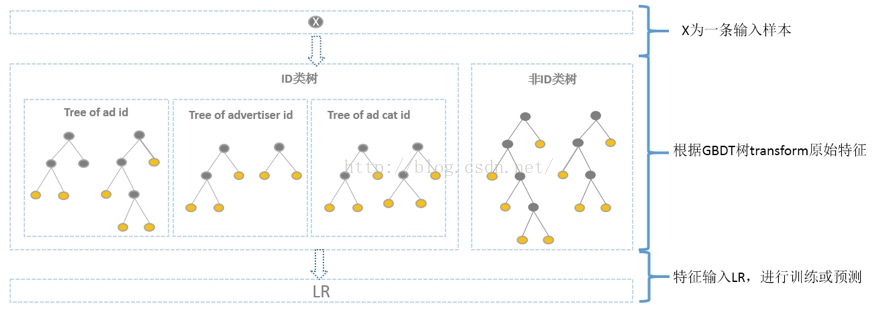
Shrinkage(缩减)的思想认为：每次走一小步逐渐逼近结果，要比每次迈一大步很快逼近结果的方式更容易避免过拟合。即它不完全信任每一个棵残差树，它认为每棵树只学到了结果的一小部分，累加的时候只累加一小部分，通过多学几棵树弥补不足。数学公式描述如下：

这里v是学习速率，一般都比较小，导致各个树的残差是渐变的而不是陡变的。

## 2.4 GBDT和LR结合用于CTR预测

 如前文所述，逻辑回归（LR）模型是一种很好的分类器，很适合广告点击预估的应用场景，然而缺点是不易选择有效的特征。GBDT的思想使其具有天然优势可以发现多种有区分性的特征以及特征组合，决策树的路径可以直接作为LR输入特征使用，省去了人工寻找特征、特征组合的步骤。这种通过GBDT生成LR特征的方式（GBDT+LR）最早由Facebook提出[15]，应用效果不亚于人工特征选取。具体过程如图所示：

上图中x为一条输入样本，遍历两棵树后，x样本分别落到两颗树的叶子节点上，每个叶子节点对应LR一维特征，通过遍历树就能得到该样本对应的所有LR特征。由于树的每条路径是通过最小化均方差等方法最终分割出来的有区分性路径，因此根据该路径得到的特征和特征组合都具有相对区分性。

 具体应用到广告点击预测中，常用的做法是：对于曝光充分训练样本充足的广告，单独建树，称为ID类树，用于发掘对单个广告有区分度的特征，但对于曝光不充分样本不充足的长尾广告，不以细粒度的属性建树，此类树作为非ID类树，即便曝光少的广告、广告主仍可以通过此类树得到有区分性的特征和特征组合。当一条样本进来时遍历两类树到叶子节点，得到的特征作为LR的输入。当广告曝光不充分不足以训练树时，其它树恰好作为补充。过程如下图所示：

通过GBDT映射得到的特征空间维度等于GBDT树的叶子节点数。一个叶子节点对应一种有区分性的特征或特征组合。针对不同的应用场景可以选择不同的GBDT建树方案，以达到最优效果。

## 2.5 在线广告点击预测系统构成

一个完整的可商业化运行的在线广告点击预测系统主要包括：1)请求接口：一般多用基于HTTP的web服务器；2)计算服务，为了处理大规模高并发高响应请求，一般使用分布式计算集群；3)数据存储服务，用于存储访问用户的特征信息以便更新模型用。一些大型企业一般会根据自身需求专门开发相关的系统，这样可以针对具体业务做专门的优化，而大多数公司，尤其是一些初创公司，通常选用开源的工具搭建自己的平台，以达到节省成本，快速开发部署，将主要精力投入到具体业务上来。本文采用后一种方法。下面分别介绍搭建在线广告点击预测系统所用到的工具。

### 2.5.1 HTTP服务器Ngnix

由于广告点击率预测系统有高并发、低延迟的性能要求，因此在选择前端处理HTTP请求的web服务器上，ngnix(<http://www.nginx.org>)成为这类系统的首选。Nginx是一款开源服务器软件，兼有HTTP服务器和反向代理服务器的功能。其主要特点在于高性能、高并发和低内存消耗，并且具有负载均衡、高速缓存、访问控制、带宽控制以及高效整合各种应用的能力，这些特性使得Nginx 非常适合计算广告这种并发很高的互联网服务。此外，Nginx 还提供了fastCGI这一与各种编程语言之间的通信接口，开发者可以很方便地将服务器的功能逻辑用fastCGI插件的形式实现，而无需关注响应HTTP 请求的细节。在广告点击预测系统中，用Nginx作为前端Web服务器，而将广告投放机的功能用C/C++语言实现成fastCGI插件，是一个开发成本较低、性能又很不错的方案。实际上，这一方案已经实现了一个基本的广告投放机，从事最简单的广告投放业务，而其他模块和功能则可以根据需求逐步开发[28]。

### 2.5.2 分布式配置和集群管理工具ZooKeeper

实际商业运营中的广告点击预测系统的流量惊人，单台服务器往往不能满足需要。在使用多台服务器的时候，会遇到很多诸如配置文件更新、集群上下线管理等分布式环境下的同步问题。ZooKeeper(http://zookeeper.apache.org)是解决这些问题非常有用的开源工具。

ZooKeeper 是为分布式应用建立更高层次的同步、配置管理、群组以及名称服务的通用工具。它的基础原理是参考文献[29]中的Paxos算法，在编程上ZooKeeper的设计很简单。所使用的数据模型非常类似于文件系统的目录树结构和Windows中注册表的结构，有名称、树节点、键/值对等，可以看作一个树形结构的数据库，可以分布在不同的机器上做名称管理。由于ZooKeeper并不传递计算数据而是传递节点的运行状态，所以运行负载很低。ZooKeeper另一个典型应用是：当某台服务器宕机或者新机器上线，Nginx的负载均衡方案需要及时作出调整。显然，如果用人工地维护，响应时间会很长，不可避免地会带来一些流量上的损失。利用ZooKeeper的Ephemeral类型节点可以很方便地实现此功能[30]。

### 2.5.3 全文检索引擎Lucene

大多数广告业务在初始运营阶段并不见得需要一个真正的倒排检索引擎，不过当广告业务开始面向长尾广告主，广告库规模较大时，采用“倒排检索”加“排序”这样的两段式决策过程是必要的。然而，实现一个功能全面、效率较高的倒排索引并不是一件简单的事，并且由于其与核心业务逻辑关系并不大，也可以用开源方案来实现。在开源工具中，Lucene(http://lucene.apache.org)是比较常用的基于Java的全文检索工具包。Lucene并不是一个完整的搜索引擎，但是针对计算广告系统的需要，它可以方便地实现全文索引和检索功能。Lucene能够为文本类型的数据建立索引，其主要功能是替文档中的每个关键词建立索引[31]。另外，Lucene还提供一组解读、过滤、分析文档，编排和使用索引的API。之所以选用Lucene除了它的高效和简单外，还因为它允许用户对其中的关键环节自定义功能逻辑。不过一些特殊的检索算法，如相关性检索，在Lucene中并不能直接支持，还需要在深入理解源代码的基础上改动或者另行开发。

### 2.5.4 跨语言通信接口Thrift

如前文所述，一个完整的在线广告点击预测系统由若干不同的模块组成，这些模块可能是用不同的编程语言开发的。为了方便在不同语言的模块之间实现调用接口，避免应用开发者过多地将精力放在底层通信上，跨语言通信协议就是用来解决这个问题。其中比较著名的有Apache Thrift，gRPC等。以本系统所使用的Thrift (http://thrift.apache.org)为例，它有自己的跨机器的通信框架，提供了一套代码生成工具，可以生成多种编程语言的通信过程代码，目前支持开发用的大多数编程语言如C++、Java、Python、PHP、Ruby、Erlang、Perl、Haskell、C#、Cocoa、JavaScript等。Thrift有一种描述对象和服务的界面定义语言IDL(Interface Definition Language)，它提供了一种网络协议，使用这些对象和服务定义的进程之间基于这种网络协议彼此进行通信，因此服务器端实现语言不会影响到客户端，这给系统各模块之间的通信提供了很大的便利[32]。此外，Thrift 还提供了实践中非常有用的版本兼容性功能，即服务器端能在不影响现有的客户端的情况下增加数据结构、字段、服务方法和函数参数。这一特性使得大型工程中模块间的依赖性大为减弱，也能够显著降低开发成本。

### 2.5.5 数据高速公路Flume

像广告点击实时预测这样的高并发系统，其在日常运行中产生的日志量也非常大。为了不影响效率应该避免对数据做单点的集中式读写，而是尽量应该让数据的处理形成环形的流动，即由数据高速公路将线上日志准实时地送至离线或在线处理平台，再将处理结果存放在缓存中供线上决策使用。在这样的架构中，一个分布式、高吞吐率的数据传送通道至关重要。

在这类数据传输工具中，Flume(http://flume.apache.org)是比较常用的开源解决方案之一。Flume是Cloudera提供的一个高可用的、高可靠的、分布式的海量日志采集、聚合和传输的系统，它支持在日志系统中定制各类数据发送方，用于分布式地收集和汇总日志数据。Flume提供了从控制台、RPC（thrift）、文本、Tail操作（UNIX tail）、日志系统（syslog，支持TCP和UDP两种模式）以及命令执行（exec）等数据源上收集数据的能力。同时，Flume还提供了对数据进行简单处理并输出到各种数据接收方的能力。如果广告投放机采用syslog方式记录投放、点击等日志，可以很方便地通过配置Flume将日志传送到Hadoop上。

### 2.5.6 分布式数据处理平台Hadoop

在线广告点击预测系统在日常运行中会生成大量的离线数据，离线数据处理部分需要一个能够存储和加工海量数据的基础设施，实际上这也是大多数大数据系统都需要的平台。在开源的解决方案中Hadoop（http://hadoop.apache.org）几乎是工业界的标准选择。Hadoop的核心架构主要包括分布式文件系统HDFS、MapReduce和HBase，其中HDFS是GFS(Google File System)[33]的开源实现，MapReduce是Google MapReduce[34]的开源实现，而HBase则是Google BigTable[35]的开源实现。

Hadoop MapReduce是一种分布式计算框架，它由两个部分组成：Map和Reduce。Map是将一个作业分解成多个任务，而Reduce是将分解后多任务处理的结果汇总起来。在程序设计中，一项工作往往可以被拆分成为多个任务，任务之间的关系可以分为两种：一种是不相关的任务，可以并行执行；另一种是任务之间有相互依赖，先后顺序不能够颠倒，这种任务是无法并行处理的。MapReduce适用于第一种类型，庞大的集群可以看作是硬件资源池，将任务并行拆分，然后交由每一个空闲硬件资源去处理，能够极大地提高计算效率，同时这种资源无关性对于计算集群的横向扩展提供了最好的设计保证。为了降低MapReduce编程的复杂性，人们还开发了Hive、Pig等开源工具产品，使用类似于SQL的脚本语言发起各种数据计算任务。

在广告系统中，Hadoop主要承担着离线数据的存储和计算需求，是系统进行大规模数据处理不可或缺的基础平台。无论是受众定向、点击率预测还是基础的报表生成，都需要在Hadoop上进行大规模的数据处理。

### 2.5.7 特征在线缓存Redis

在广告点击预测系统中，需要缓存点击率模型的参数和特征数据，以便用作模型定期更新。由于数据规模比较大，一般来说都无法直接存放在在线广告投放机的内存中，而是要用独立的缓存服务。在线用到的特征缓存有两个显著的特点，首先是往往只需要存储简单的键/值对，其次是大多数情形下需要支持高并发的随机读和不太频繁的批量写。在这样的需求下， Redis（http://www.redis.io）是比较合适的开源工具之一。

Redis是一种NoSQL数据库，它主要提供的是高性能的键值存储，采用的是内存数据集的方式。Redis的键值可以包括字符串、哈希、列表、集合和有序集合等数据类型，因此也被称作是一款数据结构服务器。Redis会周期性地把更新的数据写入磁盘或者把修改操作写入追加的记录文件，并且在此基础上实现了主从同步，具有非常快速的非阻塞首次同步、网络断开自动重连等功能。同时，Redis还具有其他一些特性，其中包括简单的check-and-set机制、pub/sub和配置设置等，使得它能够表现得更像高速缓存。Redis 还提供了丰富的客户端，支持现阶段流行的大多数编程语言，使用起来比较方便。在实际应用中当以批处理方式更新其中内容时，为了避免对线上高并发的读请求产生影响，一般采用多次写入的方案。

## 2.6 本章小结

本章主要介绍了广告点击行为预测的主要的算法逻辑回归模型与实现Ftrl；特征工程方法GBDT决策树；以及搭建一个可商业化运行平台用到的主要开源工具，这些算法和工具是构建本文的在线广告点击预测系统的基础设施。

# 第三章 系统需求分析

在线广告点击行为预测作为广告投放的一个重要环节，其流程涉及一系列复杂的数据处理与计算，根据业务类型和实现方法的不同，自然有其特殊的需求。

## 3.1 性能需求

作为一个可商用的在线广告点击预测系统，在性能上不仅要求有很高的响应速度，而且还要能够及时地更新在线预测所使用的模型，保证使用的是最新最近的数据。此外对预测结果的准确度也是越高越好，因为这将直接影响广告收益。

在响应速度方面，为了能够达到毫秒级的响应速度，除了采取线下训练，线上预测并更新模型的方法之外，本文专门为这个在线广告点击预测系统实现了一个基于MPI的分布式计算集群，将收到的请求按照负载均衡的原则平均分配到集群中的每台主机上，借此提高并发度，加快响应速度。分布式集群架构如图所示：

其中APIServer是一个http服务器，负责接收请求数据，然后把收到的请求平均分配到计算集群上，集群中的每台主机都有一个统一的算法（LR）模型，计算主机要做的工作是：1）请求数据预处理，主要是特征提取和转换；2）用现有的LR模型结合请求数据特征做出预测，即用户是否会点击该广告；3）将特征数据缓存到Redis服务器上，并待用户实际行为（点击动作是否发生）后更新特征缓存中的记录。模型在线训练更新服务会定期用Redis特征缓存里积累的数据对线上模型做增量训练更新，然后分发给每台计算主机用作后续预测。

## 3.2 功能需求

本系统是基于用户微博数据的在线广告点击预测系统，除了具有一般CTR预估系统的功能如模型训练与更新之外，还需要有必要的对微博数据（主要是文本数据）进行处理的能力，如分词、词性标识、关键词提取、主题预测、文章向量化表示等。而这些功能主要是基于jieba实现的。

除了线上的功能之外，线下的功能需求主要是文本的特征工程，即将用户的微博文章数据转换为能被算法（逻辑回归）模型所处理的数值向量。这个过程主要由word2vec、article2vec、xgboost等工具实现，为了提高效率，在本系统的分布式集群计算服务中也添加了对线下功能服务的支持，具体过程在下一章详细论述。

## 3.3 非功能需求

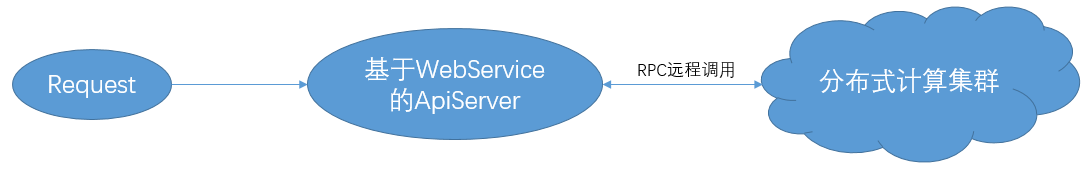
除了支持业务的功能需求和性能需求之外，非功能需求主要是易用性和可扩展性。

易用性方面：要求分布式集群上的所有算法通过APIServer对外提供两种服务接口，一是基于WebService的实时线上请求，主要用于提供在线服务的线上算法；二是命令行请求，用于批处理数据文件，主要用于线下算法如文本处理、分词、特征工程。

可扩展性方面：要求系统提供二次开发接口，一个现成的算法只要按照系统规范要求进行必要的改造和封装，就能够无缝兼容与系统，提供线上服务。添加算法无需修改或重新编译现有模块。

# 第四章 系统设计与实现

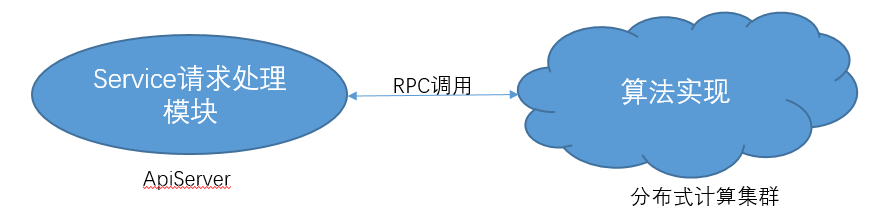
## 4.1 系统整体架构

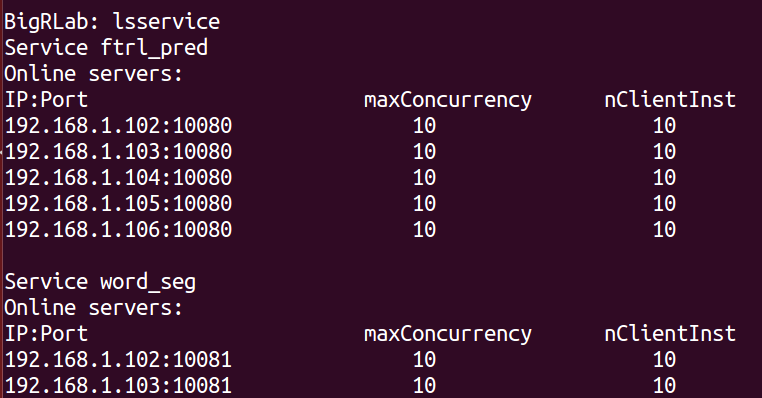
 为了满足性能上高并发、快响应的需求，系统整体采用一个请求处理/分发单元+多个计算单元组成的分布式计算集群的总-分架构。如图所示：

ApiServer是一个基于WebService的在线请求处理服务器，用于接收外部请求，并将收到的请求转换为RPC调用并按照负载均衡的原则分发到计算集群上的主机。计算集群完成请求的计算任务并将结果返回给ApiServer，ApiServer将收到的结果以WebService的形式（http+json）返回给请求者。

分布式计算集群上除了算法运行任务的计算服务器之外，还包括用于特征缓存的数据服务器和用于模型更新的服务器。特征缓存服务用作当收到广告点击预测请求时缓存请求数据中的用户特征和用户实际行为（点击或未点击），当缓存到一定数目时通过增量训练更新线上模型，从而能够实时反映用户兴趣行为的变化。

## 4.2 Service定义

 在本系统中定义一个service为能够对外提供线上服务的算法运行实例，主要包括请求处理模块（ApiServer端运行）和算法实现模块（在分布式集群上运行），

算法实现模块主要由运行算法的计算服务主机构成，根据业务需求可能也会有其他的支持服务如特征缓存数据库等。在系统控制台端可以通过lsservice命令查看当前线上运行的service列表，如图所示：

可以看到service的服务器列表，每个服务器属性主要有IP地址和端口号、支持的最大并发数（Thrift RPC Server参数）、以及当前的连接数（从ApiServer发起的连接）。

 由于不同的业务处理请求的方式不同，当向系统中添加一个新的service时，除了要实现分布式集群上运行的算法模块之外，还要实现ApiServer上运行的请求处理模块，各个service请求处理模块由service的二次开发人员所编写，因此和ApiServer是松耦合的，支持运行时动态添加而不必重新编译ApiServer。

在Service基类中提供两个纯虚函数，分别是handleRequest和handleCommand，前者用来处理http在线请求，对外提供WebService服务，后者提供基于命令行的用户接口。每一个具体的service派生类都要根据自己特定的业务需求和处理方法提供这两个函数的功能实现。

此外，ApiServer端运行的Service请求处理模块还负责计算集群上服务器的管理，包括新服务器加入，已有服务器下线，以及服务器异常处理；还负责将服务请求按照一定的规则分发给计算服务集群上的服务器处理。系统默认的是基于负载均衡的分发方案，随机挑选一个服务器处理当前请求，各具体的Service派生类可根据自身业务需求改写这一方案。对于异常处理：如集群中的某一台服务器崩溃，则RPC调用会抛出异常或请求超时，此时系统会自动重新选择另一台服务器处理请求，并将出现问题的服务器从候选列表中删除。

## 4.3 ApiServer的设计与实现

ApiServer在本系统中主要负责接收外部业务处理请求并且请求分发到计算集群上处理，其主要包括3个模块，Service管理，请求处理，和算法服务集群管理，本节将分别对这3个模块展开介绍。

### 4.3.1 Service管理模块

在本系统中ApiServer主要任务是接收外部请求，并交给请求所指定的service处理。为了能够支持运行时动态加载，所有在ApiServer上运行的service请求处理模块都被要求编译成动态链接库。在ApiServer端通过loadlib命令加载。为了省去每次添加新的service都要手动加载的麻烦，可以将service动态链接库路径记录在autoload.conf文件中，每次ApiServer启动时会先依据此文件记载的内容加载相应的service请求处理动态链接库。由加载的service动态链接库创建service实例。

Service动态链接库和service实例都由ServiceManager类管理，该类采用单例模式，只有一个全局唯一实例。ServiceLib类是对service动态链接库的类抽象，每一个该类的对象拥有一个service动态链接库的handle，此外提供两个函数用于创建该service实例和获取库的名称；每一个service动态链接库都有一个全局唯一的名称，此外还必须提供用于创建实例和获取名称的函数接口以供ServiceLib调用。它们之间的关系如图所示：



### 4.3.2 请求处理模块

ApiServer接受的外部请求分为两种，一种是WebService请求，通过http协议传输json封装的字符信息，用于处理实时在线请求，每次请求只处理一条数据并立刻返回结果；另一种是终端输入的命令行，用于线下批量处理存放于文件中的请求数据，结果一般输出到指定的文件中。

对于WebService请求，请求发送者将处理该请求的service名称封装在URL中，一般格式为：

<http://apiserver_addr:port/service_name>

当ApiServer的http服务器模块收到该URL请求后，会通过service\_name字段得到请求服务的service名称，然后通过ServiceManager查找当前系统中有没有该service，若没有返回错误信息，若找到则将请求正文内容作为参数传给该service的handleRequest函数处理。处理结果同样以json格式的文本返回给请求者。

对于命令行格式的请求，其一般格式为：

service $service\_name $args

ApiServer的命令行处理模块从终端读入一行用户输入，若第一个单词为service则表示该命令请求要交给具体的service处理，否则都是系统命令；紧接着读入指定处理该请求的service名称service\_name，然后通过ServiceManager查找这个service，若找到就将剩余的args参数列表传给该service的handleCommand处理，若找不到指定的service就在终端上输出错误信息。



### 4.3.3 算法服务管理模块

 在计算服务集群上，每一个算法实例启动时都要连接ApiServer（由启动参数指定ApiServer的地址和端口）以完成注册，这样集群上的该算法实例会成为对应service的计算集群的一员，成为一个作业分配的候选者。这个过程在ApiServer端是由算法服务管理模块AlgManager实现的。AlgManager本身也是一个Thrift RPC Server，计算集群上的算法实例通过Thrift RPC远程过程调用和ApiServer进行通信。AlgManager提供如下接口：

当计算集群上算法实例启动时会将自己的信息填入到AlgSvrInfo结构中，然后以此为参数远程调用ApiServer上AlgManager模块的addServer完成注册；在该算法实例退出的时候相应地会调用rmServer从ApiServer上删除自己的注册信息。AlgSvrInfo记录了一个算法实例的关键信息，定义如下：

其中addr和port分别定义了该算法实例所在集群主机的ip地址和所用的服务端口号；maxConcurrency指定了该算法实例所支的最大并发数，即同时能处理的最大请求数；serviceName指定了该算法实例属于哪一个service；libName指定了ApiServer所加载的service请求处理动态链接库的名称标识。

 计算服务集群上一个新的算法实例加入系统过程如图所示：

 当计算集群上一个算法实例退出时会远程调用AlgManager的rmServer，进而从对应service服务器列表中删除自己和已经建立的连接，如果删除后该service服务器列表为空，那么ServiceManager会删除该service的信息，后续到来的指定该service处理的请求都会返回错误。过程如图所示：

### 4.3.4 负载均衡和异常处理

为了最大限度提高并发处理能力，对于多台主机构成的计算集群，负载均衡是必须要考虑的。ApiServer在service基类提供了一套默认的负载均衡方案，该方案为每个service创建一个空闲RPC连接池，当新的算法服务器加入的时候，根据其声明的maxConcurrency创建相应数目的空闲连接，然后放入空闲连接池中并执行random permutation的随机乱序操作，确保连续到来的请求尽可能地平均分配到不同的主机上。当一个请求到来的时候，该默认调度方案会从空闲连接池中取出第一个空闲连接，用它完成指定的RPC调用请求后再放回到空闲连接池中以供后续请求使用。每一个继承自service的具体派生类（动态链接库实现的）都可以定义自己的调度方案以覆盖默认方案，只需重写addServer函数并提供相关的数据结构支持即可。

异常处理方面，ApiServer端将异常分为用户异常和系统异常。用户异常是由算法定义的逻辑错误如非法输入，此类异常定义为InvalidRequest异常类，是Thrift Exception的子类。当ApiServer执行远程调用请求时收到这类异常会直接返回给调用者错误信息（终端输出或json文本，取决于请求类型是命令行还是WebService），并按正常流程将用过的RPC连接放回到空闲连接池中；系统异常主要是不可恢复的错误，如远程服务器意外停机，此时若在已经意外断开的连接上执行RPC调用会抛出Thrift Exception异常，对于这类异常系统不会将当前使用的RPC连接放回连接池。为了及时更新各线上算法计算服务器的状态，要求集群上的算法服务器每隔15秒钟向ApiServer发送一次信号，由AlgManager的informAlive远程函数实现，若连续1分钟未收到该主机的信号，系统会从相应service的服务器列表中将其删除。

## 4.4 算法服务的设计与实现

算法服务是对现有算法的改造封装以便和本系统兼容，算法服务实例运行于分布式计算集群上，通过Apache Thrift RPC远程调用和ApiServer端的service lib通信。算法服务本身是一个Thrift Server，按照自身业务逻辑提供远程调用接口供ApiServer端的service lib使用，计算集群上的算法服务实例和ApiServer端该算法对应的service lib实现共同构成了一个完整的service服务。

当算法实例启动时，要填写一个AlgSvrInfo对象，包括自己所在主机的ip地址，和本实例的服务端口号；ApiServer端service lib的名字；service名称；以及所支持的最大并发数，使用这个AlgSvrInfo对象去ApiServer注册，ApiServer会根据service名称找对应的service，若找到就把这个新到的实例加入到其服务器列表中，若找不到就用service lib名称找对应的动态链接库句柄，并用这个链接库创建一个新的service。过程如图所示。

为了防止ApiServer意外崩溃，算法实例每隔一段时间（15秒）会启动一次注册流程，ApiServer收到注册请求若发现该实例已经注册，会直接返回实例已注册的错误，算法实例这边会忽略这个错误。这样当ApiServer崩溃重启后，之前所有在线的算法实例都会自动重新上线而不需要人工手动重启。

此外，为了实现模型自动更新，算法实例会自动保存启动时指定的模型数据文件路径，之后若要更新模型文件，只需把要更新的文件放在源文件的同一目录下，并在源文件名之后加上.update后缀，如源文件是model.dat，更新文件就是model.dat.update，算法实例会每隔一段时间检查是否有新的模型更新，若发现更新，就自动下线服务，然后将模型更新文件替换掉源文件，再初始化算法，到ApiServer注册，完成重新上线服务。

## 4.5 本章小结

本章详细介绍了面向微博的在线广告点击预测系统的设计和实现，首先介绍了系统整体架构和系统中的核心概念service，然后详细介绍了ApiServer各个组件和他们之间的关系以及与系统其他模块间交互流程，最后介绍了分布式计算集群上算法服务模块的设计和实现。

# 第五章 业务流程

## 5.1 微博文章处理

本系统使用用户感兴趣的微博文章（包括原创、转发、点赞等操作）作为用户的非结构化特征。首先要做的是将这些文本数据转换成可供机器学习算法使用的数值型数据。本系统使用jieba分词和word2vec来实现。下面分别予以介绍。

### 5.1.1 Jieba分词

对文章数据的第一步操作是合并分词。合并就是将用户感兴趣的文章连接成一篇长文章，然后用jieba进行分词并且依据事先定义的停用词列表去掉停用词。为后续的文章向量化作准备。Jieba是很好的中文自然语言处理工具[37]，其分词功能十分强大，采用了基于前缀词典实现高效的词图扫描，生成句子中汉字所有可能成词情况所构成的有向无环图；并用动态规划查找最大概率路径, 找出基于词频的最大切分组合；对于未登录词，采用了基于汉字成词能力的HMM模型结合Viterbi算法处理。Jieba有3种分词模式，分别是：1）精确模式，试图将句子最精确地切开，适合文本分析；2）全模式，把句子中所有的可以成词的词语都扫描出来, 速度非常快，但是不能解决歧义；3）搜索引擎模式，在精确模式的基础上，对长词再次切分，提高召回率，适合用于搜索引擎分词。此外还支持用户自定义词典，可以根据具体应用环境使用不同的词典，如医学、IT、时尚等等。同时Jieba具有并行分词能力，能极大提高分词效率。关键词抽取方面，实现了基于TF-IDF和TextRank算法[36]的关键词提取并有很好的效果。总之，Jieba是中文自然语言处理的多面手工具，其高效的处理速度和高精确度为后续的特征转换和算法处理提供可可靠的保证。

### 5.1.2 文章向量化

接下来是将分好词的用户感兴趣的文章集合转换成数值向量。首先是用word2vec工具建立每个词的向量。Word2vec 是 Google 在 2013 年推出的一款将词表征为实数值向量的开源高效工具[38]，其利用深度学习的思想，通过训练把对文本内容的处理简化为 K 维向量空间中的向量运算，而向量空间上的相似度可以用来表示文本语义上的相似度。Word2vec输出的词向量可以被用来做很多 NLP 相关的工作，比如聚类、找同义词、词性分析等。Word2vec 使用的是 Distributed representation 的词向量表示方式，其基本思想是通过训练将每个词映射成 K 维实数向量（K 一般为模型中的超参数），通过词之间的距离（比如 cosine 相似度、欧氏距离等）来判断它们之间的语义相似度。其采用一个三层的神经网络，输入层-隐藏层-输出层。根据词频的Huffman编码，使得所有词频相似的词在隐藏层激活的内容基本一致，出现频率越高的词语，他们激活的隐藏层数目越少，这样有效的降低了计算的复杂度。这个三层神经网络本身是对语言模型进行建模，但也同时获得一种单词在向量空间上的表示，而这正是Word2vec的真正目标。Word2vec大受欢迎的一个原因是其高效性，一个优化的单机版本一天可训练上千亿词。

Word2vec有两种输出形式，一种是词的向量表示，此时要用-size参数指定向量的维度，维度越高可利用的特征越多，分类越准确，但计算开销也会增大，速度随之会变慢；另一种是类标识classid，需要事先用-nclasses参数指定类的个数，同一类里的单词有较高的相似度。相应地将文章转换成向量也有两种方法，一种是将文章中每一个词对应的向量相加最后求平均得到文章的向量表示；另一种是初始化nclasses维向量，元素全为0，遍历文章中每一个词，找到其classid，并将向量中下标是classid的元素加1，最后对向量进行归一化得到文章的向量表示。实践发现后一种方法有较好的准确度。

## 5.2 GBDT特征工程

在得到了用户感兴趣的文章向量表示后还不能直接用于ctr预估，因为文章的向量表示是浮点实数型向量，无法与已有的结构化特征结合。需要转换成与结构化特征相同的形式。这时候用到第2.3,2.4节所介绍的GBDT方法做特征转换，又称特征工程。

本系统选用xgboost实现GBDT特征工程，xgboost全称是eXtreme Gradient Boosting，是Gradient Boosting Machine的一个C++实现，其目标是提供一个可扩展、可移植、分布式的Gradient Boosting算法库；它支持常见的分布式框架如Hadoop、Spark；并且提供多种语言的编程接口如C++、Java、Python；具有很高的运行效率和准确度。

Xgboost做GBDT特征工程过程如下：首先由用户感兴趣的文章向量和最终期望输出（用户实际点击广告行为，点击为1，未点击为0）构成训练样本，用xgboost训练得到一个模型，而模型是由若干颗决策树组成。对于新到来的数据（文章向量），用此模型跑一编预测过程，这里我们并不关心预测结果，而是使用模型中每颗决策树决策路径终点叶子的序号为新的特征，这些新的非结构化特征是离散的，与用户原有的结构化特征形式一致。将这两种特征拼接起来（新得到的特征起始序号在原有结构化特征的最大序号上加1）就得到了最终送给lr预估算法做预测的用户特征数据。特征工程的任务到此结束。

## 5.3 LR预估

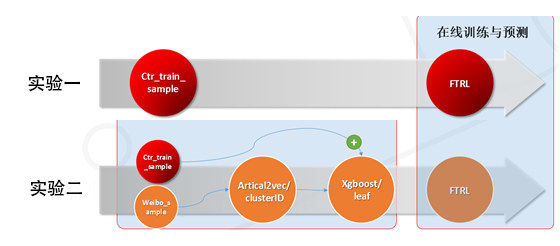
在第二章中提到，逻辑回归模型是一种很高效的分类器，很适合做CTR预估。本系统中使用ftrl\_proximal\_lr[39]实现这一功能。ftrl\_proximal\_lr实现了ftrl逻辑回归算法，并且最大限度地利用并行多线程以提高效率；而且支持模型在线增量更新，即在已有模型的基础上用线上新获取的数据（用户的特征和具体点击行为）来增量训练，矫正原有模型的各个参数，从而实现更高的预测准确度。通过本系统的线上运行实验证明ftrl\_proximal\_lr有很高的运行效率和准确度，能够应对实际生产环境中的高并发请求。

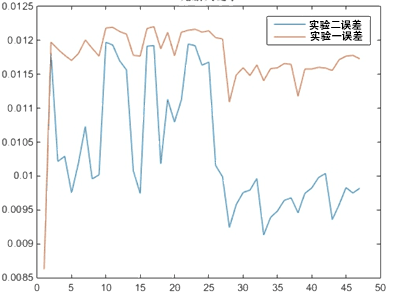
## 5.4 本章小结

本章主要介绍了系统的业务处理流程，先将用户感兴趣的微博文章通过特征工程转换为和原有的结构化特征形式一致的离散型数值特征，再和原有结构化特征结合形成新的组合特征，最后交给ftrl算法做预估。

# 第六章 系统测试

## 6.1 线下测试

 为了测试本文ctr预估方法的性能，分别设计了两组组实验加以比较。实验过程如图所示：

其中实验一是直接用ftrl算法在用户的结构化特征数据上做训练和预测；实验二是用本文的方法将用户的微博行为数据通过特征工程转化为数值特征，再和用户原有的结构化特征结合生成新数据，在之基础上用ftrl做训练和预测。实验结果如下图所示：

图中橙色线条表示用传统方法实验一的误差率，蓝色线条表示用本文的方法实验二的误差率。可以看出本文所介绍的方法能够显著地降低误差。

## 6.2 线上环境测试

线上运行主要考虑并发量和相应时间两个性能指标。由于本系统使用了分布式计算集群加快处理速度，因此在相应时间和吞吐量上都能达到较为理想的效果。（待补充）

# 第七章 总结与展望

## 7.1 总结

本文旨在介绍一种结合用户微博行为记录的广告点击率预测方法，并实现了一个可商用的基于此种方法的在线广告点击预测系统。本文所用方法的重点是特征工程，即如何自动选取有用特征。为此首先将用户感兴趣的微博文章转换为数值向量，用这些向量和用户最终点击行为数据训练出一个迭代决策树GBDT模型，再用此GBDT模型将连续的微博文章向量特征转换成和结构化特征形式一致的离散特征，实际上是使用了GBDT模型中每棵树不同的叶子序号作为特征。

系统实现方面，本文考察了多种搭建互联网大数据处理平台的开源框架，综合利用多种开源工具，设计并实现了一个分布式大数据处理平台，其主要优势在于将大量计算任务分散到多台服务器上执行，即能够处理线上实时请求，也能够处理线下作业，同时有很好的扩展性和兼容性。经过二次开发也可用于处理其他类型的业务，具有很好的通用性。

## 7.2 展望

本文提出了一种基于用户微博的广告点击预测，其目的是通过用户微博浏览记录发现用户兴趣，从而更准确地投放广告。整个业务流程中，微博文章数据的特征转换是关键环节。目前所用的方法是将用户感兴趣的所有文章拼接成一篇长文章，再将拼接后的文章转换为数值向量，转换方法是用每个词的向量叠加并归一化。这个过程没能够考虑到不同的文章所代表的不同的主题，而拼接过程事实上是消除了不同文章间的个性差异。势必会对最终预测准确度造成负面影响。在这方面今后的改进方向是使用NLP技术如话题模型找出每篇用户感兴趣文章的话题关键词，给这些关键词更高的权重，并以此为基础分析用户兴趣以达到更高的投放准确度。

在系统实现方面，为了支撑业务需求，本文实现了一个基于分布式的大数据处理平台，但目前这个平台功能上还有很多待完善的地方，目前线上只支持预测（支持在线更新模型的算法如Online Ftrl除外），训练模型的过程需要线下完成。同时还要考虑加入用户权限管理功能，目前是完全开放的。最终目标是将其打造成一个功能完善的，支持多用户、多任务的综合大数据处理平台。

# 参考文献

1. 第38次中国互联网络发展状况统计报告 <http://www.cnnic.net.cn/>
2. 新浪财经频道 <http://finance.sina.com.cn/roll/2016-11-12/doc-ifxxsfip4572557.shtml>
3. 刘鹏, 王超, 计算广告,互联网商业变现的市场与技术, 人民邮电出版社, 2015
4. Zhang, Wei Vivian, and Rosie Jones. "Comparing click logs and editorial labels for training query rewriting." WWW 2007 Workshop on Query Log Analysis: Social And Technological Challenges. 2007.
5. Richardson M, Dominowska E, Ragno R. Predicting clicks: Estimating the click-through rate for new ads, Proceedings of the 16th International conference on World Wide Web, Banff, Canada, 2007: 521-530
6. Guo F, Liu C, Kannan A, et al. Click chain model in web search, Proceedings of the 18th International conference on World Wide Web. Madrid, Spain, 2009: 11-20
7. Jahrer M, Toscher A, Lee J Y, et al. Ensemble of collaborative filtering and feature engineered models for click through rate prediction, Proceedings of the 18th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDDCup Workshop, Beijing, China, 2012
8. Shen S, Hu B, Chen W, et al. Personalized click model through collaborative filtering, Proceedings of the 5th ACM International Conference on Web Search and Data Mining, Seattle, USA. 2012: 323-332
9. Regelson M, Fain D. Predicating click-through rate using keyword clusters, Proceedings of the second workshop on Sponsored Search Auctions, Ann Arbor, USA, 2006: 9623-9628
10. Agarwal, Deepak, et al. "Estimating rates of rare events at multiple resolutions." Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2007.
11. Wang X, Li W, Cui Y, et al. Click-through rate estimation for rare event in online advertising, Hua Xian-Sheng, Mei Tao, Hanjialic A eds. Online Multimedia Advertising: Techniques and Technologies. Hershey Pennsylvania, USA: IGI Global, 2010: 1-12
12. Kempe D, Mahdian M, A cascade model for externalities in sponsored search, Proceedings of the 4th International Workshop on Internet and Network Economics, Chicago, USA, 2008: 585-596
13. Guo F, Liu C, Wang Y M, Efficient multiple-click models in web search, Proceedings of the 2nd ACM International Conference on Web Search and Data Mining, Barcelona, Spain, 2009: 124-131
14. 腾讯大数据：CTR预估中GBDT与LR融合方案 <http://www.cbdio.com/BigData/2015-08/27/content_3750170.htm>
15. He, Xinran, et al. "Practical lessons from predicting clicks on ads at facebook." Proceedings of the Eighth International Workshop on Data Mining for Online Advertising. ACM, 2014.

1. <https://github.com/guestwalk/Kaggle-2014-criteo>
2. Cox, DR. "The regression analysis of binary sequences (with discussion)". J Roy Stat Soc B. 20, 1958: 215–242.
3. Walker SH, Duncan DB. "Estimation of the probability of an event as a function of several independent variables". Biometrika. 54, 1967: 167–178.
4. <https://en.wikipedia.org/wiki/Logistic_regression#Logistic_function.2C_odds.2C_odds_ratio.2C_and_logit>
5. Hosmer, David W.; Lemeshow, Stanley. Applied Logistic Regression (2nd ed.). Wiley. ISBN 0-471-35632-8, 2000
6. Everitt, Brian. The Cambridge Dictionary of Statistics. Cambridge, UK New York: Cambridge University Press, 1998
7. Tjalling J. Ypma, Historical development of the Newton-Raphson method, SIAM Review 37 (4), 531–551, 1995.
8. Menard, Scott W, Applied Logistic Regression (2nd ed.). SAGE, 2002
9. H. Brendan McMahan, Follow-the-Regularized-Leader and Mirror Descent: Equivalence Theorems and L1 Regularization, Google, Inc. 2011
10. J. Duchi and Y. Singer. Efficient learning using forward-backward splitting. In Advances in Neural Information Processing Systems 22, pages 495-503. 2009
11. L. Xiao. Dual averaging method for regularized stochastic learning and online optimization. In NIPS, 2009
12. McMahan, H. Brendan, et al. "Ad click prediction: a view from the trenches." Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2013.
13. 陶辉, 深入理解Nginx: 模块开发与架构解析, 机械工业出版社, 2013
14. Lamport, Leslie. "Paxos made simple." ACM Sigact News 32.4 (2001): 18-25.
15. Junqueira, Flavio, and Benjamin Reed. ZooKeeper: distributed process coordination. " O'Reilly Media, Inc.", 2013.
16. McCandless, Michael, Erik Hatcher, and Otis Gospodnetic. "Lucene 实战." (2011): 6-7.
17. Rakowski, Krzysztof. Learning Apache Thrift. Packt Publishing Ltd, 2015.
18. Ghemawat, Sanjay, Howard Gobioff, and Shun-Tak Leung. "The Google file system." ACM SIGOPS operating systems review. Vol. 37. No. 5. ACM, 2003.
19. Dean, Jeffrey, and Sanjay Ghemawat. "MapReduce: simplified data processing on large clusters." Communications of the ACM 51.1 (2008): 107-113.
20. Chang, Fay, et al. "Bigtable: A distributed storage system for structured data." ACM Transactions on Computer Systems (TOCS) 26.2 (2008): 4.
21. Mihalcea, Rada, and Paul Tarau. "TextRank: Bringing order into texts." Association for Computational Linguistics, 2004.

1. <https://github.com/fxsjy/jieba/>
2. Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781 (2013).
3. https://github.com/bingzhengwei/ftrl\_proximal\_lr

# 致谢

本论文是在xx老师的悉心指导下完成的。xx老师作为一名优秀的、经验丰富的教师，具有丰富的xx知识和xx经验，在整个论文实验和论文写作过程中，对我进行了耐心的指导和帮助，提出严格要求，引导我不断开阔思路，为我答疑解惑，鼓励我大胆创新，使我在这一段宝贵的时光中，既增长了知识、开阔了视野、锻炼了心态，又培养了良好的实验习惯和科研精神。在此，我向我的指导老师表示最诚挚的谢意！

……

(仅为网络示例，可根据论文实际进行撰写，使用时把模板示例内容尽皆删除即可)

# 北京大学学位论文原创性声明和使用授权说明

**原创性声明**

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品或成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本声明的法律结果由本人承担。

论文作者签名： 日期： 年 月 日

**学位论文使用授权说明**

（必须装订在提交学校图书馆的印刷本）

本人完全了解北京大学关于收集、保存、使用学位论文的规定，即：

* 按照学校要求提交学位论文的印刷本和电子版本；
* 学校有权保存学位论文的印刷本和电子版，并提供目录检索与阅览服务，在校园网上提供服务；
* 学校可以采用影印、缩印、数字化或其它复制手段保存论文；
* 因某种特殊原因需要延迟发布学位论文电子版，授权学校□一年/□两年/□三年以后，在校园网上全文发布。

（保密论文在解密后遵守此规定）

论文作者签名： 导师签名：

日期： 年 月 日