第一部分 基于大数据的推荐系统开发基础

# 推荐系统导论

## 推荐系统的起源

面对大数据量的信息过载，如何从海量信息中过滤出对自己有用的信息，成为推荐系统首要解决的问题。

最早的信息过滤方法是通过类目导航的方式进行，比如早期的门户网站雅虎、新浪、搜哦等等。门户网站需要用户被动的按照设定好的类目逐层查找。后来搜索引擎出现，它只需要用户提供意图明确的查询文字就能够提供大量的相关信息。

推荐系统会根据用户以往的行为，主动提供给用户一种选择。所以，推荐系统的优势比较明显，比如用户大多数情况下并没有明确的意图。推荐系统可以帮助用户发现，带给用户惊喜。

为了解决信息过载(Information overload)的问题，人们提出了推荐系统（与搜索引擎对应，人们习惯叫推荐系统为推荐引擎）。当我们提到推荐引擎的时候，经常联想到的技术也便是搜索引擎[[1]](#footnote-1)。

二者有共同的目标，即解决信息过载问题，但具体的做法因人而异。

搜索引擎更倾向于人们有明确的目的，可以将人们对于信息的寻求转换为精确的关键字，然后交给搜索引擎最后返回给用户一系列列表，用户可以对这些返回结果进行反馈，并且是对于用户有主动意识的，但它会有马太效应的问题，即会造成越流行的东西随着搜索过程的迭代会越流行，使得那些越不流行的东西石沉大海。

而推荐引擎更倾向于人们没有明确的目的，或者说他们的目的是模糊的，通俗来讲，用户连自己都不知道他想要什么，这时候正是推荐引擎的用户之地，推荐系统通过用户的历史行为或者用户的兴趣偏好或者用户的人口统计学特征来送给推荐算法，然后推荐系统运用推荐算法来产生用户可能感兴趣的项目列表，同时用户对于搜索引擎是被动的。其中长尾理论（人们只关注曝光率高的项目，而忽略曝光率低的项目）可以很好的解释推荐系统的存在，试验表明位于长尾位置的曝光率低的项目产生的利润不低于只销售曝光率高的项目的利润。推荐系统正好可以给所有项目提供曝光的机会，以此来挖掘长尾项目的潜在利润。

如果说搜索引擎体现着马太效应的话，那么长尾理论则阐述了推荐系统所发挥的价值。

## 推荐系统的场景

推荐系统的应用

新闻资讯

视频音乐

广告购物

热度排行榜：

不同的排行榜算法：

单一维度的投票

考虑多个维度的综合打分

考虑时间因素：如引入衰减权重（半衰期、冷却定律）

考虑反馈信息：如Reddit算法

考虑置信度：如威尔逊区间

防止马太效应：如MAB算法等

评价：

容易实现，可以解决新用户的冷启动问题

更新很慢，很难推新

排行榜很多事一个业务问题，而不是一个算法问题

## 推荐系统的算法

推荐系统的核心问题，如何评估一个用户对一个物品的评分（喜欢程度）。

影响推荐效果的因素：

用户交互界面（user Interface），用户推荐系统的第一感知很重要，比如：Twitter的“favorite“图案由star换成heart

数据层面，数据收集：有效性（是否真实、准确）、全面性（是否有偏）

领域知识（Domain knowledge）：产品的定位、具体推荐需求的理解

算法迭代（Algorithm/Model）

传统的推荐系统，包括热度排行（Popularity Ranking）、协同过滤（collaborate filter）、基于内容或知识的推荐（Content-based/Knowledge-based）、混合推荐方法（Hybird Approaches）。其中最为核心的就是协同过滤，协同过滤分为Memory-Based方法和Model-Based方法，前者又可以分为Item-based和user-Based；后者有很多方法，包括频繁项挖掘、聚类、分类、回归、矩阵分解、RBM、图模型等等算法。

近年来推荐系统领域的最新研究，包括排序学习（Learning To Rank）、因子分解机（Factorization Machine）和深度学习（Deep Learning）。

协同过滤的思想来源：现实生活中朋友之间相互推荐自己喜欢的东西。基本组成元素：

N个用户，M个物品

评分矩阵：用户对物品的评分（显式）或用户对物品的行为（隐式）

相似度度量：物品与物品（用户与用户）之间的相似性

User-based CF

算法步骤

1计算目标用户的前K个相似度用户（neighborhood）

各种相似性度量

2找出相似用户喜欢的物品，并预测目标用户对这些物品的评分

预测模型：KNN，regression

3. 过滤掉目标用户已经消费过的物品

4. 将剩余物品按照预测评分排序，并返回前N个物品

item-based CF

基本思想是计算item之间两两相似性，目标用户对某个item的评分可以根据其对相似的items的评分来预测，选取用户预测评分最高的前N个商品进行推荐。

item-based在电商推荐中逐渐成为主流的原因：

降低实时计算的复杂性

解决用户的冷启动问题

user-based CF与item-based CF对比：

item-based准确性好，表现稳定可控，便于离线计算；但是推荐结果的多样性会差一点，一般不会给用户惊喜性。

user-based可以帮助用户发现新的商品，但是需要较复杂的在线计算，需要处理新用户的问题。

item之间的相似性比较单纯，是静态的；而user之间的相似性比较复杂，是动态的，应用时需要特征小心。

协同过滤的优缺点：

优点：需要很少的领域知识，模型通用性很强；工程上实现简单，效果很好

缺点：冷启动问题（new user/items），数据稀疏性问题，假设”过去的行为决定现在“没有考虑具体情景的差异

热门倾向性（Popularity Bias），很难推荐出小众偏好。

基于内容/知识的推荐

什么是内容，

文本描述：通常用NLP技术挖掘关键词

item的属性，比如电影的主题，衣服的材质等

item的特征，例如语音信号表示，图像向量表示等

基本组成

item特征向量，如文本的TF-IDF向量

用户profile向量：通常通过用户偏好的items来提取

匹配分：直接计算cosine，分了/回归模型

基于内容推荐的评价

优点

能够推荐出用户独有的小众偏好

可以一定程度解决数据稀疏问题和item的冷启动问题

通常具有较好的可解释性

不足

对于很多推荐问题，提取出有意义的特征并不容易

很难将不同item的特征组合在一起

很难带给用户惊喜性

如果用户的profile挖掘不准，推荐效果往往很差

混合方法（Hybird Approaches）

模型融合方法

加权（weighted）组合

切换（switching）：确定一个合理的切换跳进

混合（Mixed）

特征组合（feature combination）：不同模型特征组合到一起

特征扩展（feature augmentation）：一个模型的输出作为另一个的特征

级联（Cascade）：粗排 -> 精排

评价

通常比单个算法表现好

需要在不同算法之间、理论效果和实际可行性之间仔细权衡

## Learning To Rank

大多数推荐结果是一个列表，因此推荐结果的排序也很重要。

L2R是将排序看做一个机器学习问题。

排序的评价准则：

Normalized Discounted Cumulative Gain(NDCG)

Mean Reciprical Rank(MRR)

问题：很难直接利用这些指标来指导学习过程

常用算法：

Pointwise: LR, GBDT, SVM

Pairwise: RankSVM, RankBoost, RankNet, LambdaRank

Listwise: AdaRank, LambdaRank/LambdaMart, ListNet, ListMLE

# 数据形态与业务场景

## 数据形态

推荐系统一般包含输入层、模型层和输出层。输入层就是各种数据，主要包括用户的评分或者行为反馈数据、用户画像（性别、年龄、喜好等）和项目内容（文本、图像等描述或内容）数据、用户生成内容（社会化关系、标注、评论等辅助数据）。

### GroupLens电影数据

|  |  |
| --- | --- |
| 列名称 | 说明 |
| User ID | 整数类型 |
| Item ID | 整数类型 |
| 时间戳 | 行为发生的时间戳 |

MovieLen-20M示例：

评分数据集rating.csv

userId,movieId,rating,timestamp

1,2,3.5,1112486027

1,29,3.5,1112484676

1,32,3.5,1112484819

电影数据movies.csv

movieId,title,genres

1,Toy Story (1995),Adventure|Animation|Children|Comedy|Fantasy

2,Jumanji (1995),Adventure|Children|Fantasy

3,Grumpier Old Men (1995),Comedy|Romance

用户标签数据tags.csv

userId,movieId,tag,timestamp

18,4141,Mark Waters,1240597180

65,208,dark hero,1368150078

65,353,dark hero,1368150079

### 淘宝-电商数据

|  |  |
| --- | --- |
| 列名称 | 说明 |
| 用户ID | 整数类型，序列化后的用户ID |
| 商品ID | 整数类型，序列化后的商品ID |
| 商品类目ID | 整数类型，序列化后的商品所属类目ID |
| 行为类型 | 字符串，枚举类型，包括(‘pv’, ‘buy’, ‘cart’, ‘fav’) |
| 时间戳 | 行为发生的时间戳 |

UserBehavior.csv示例：

1,2268318,2520377,pv,1511544070

1,2333346,2520771,pv,1511561733

1,2576651,149192,pv,1511572885

### 线下商超数据

14列主要数据

### 数据反馈类型

根据上述数据类型的展示，可以把现有的数据根据反馈类型划分为，显示反馈类型和隐式反馈类型。显示反馈数据，用户对物品有明确的评分，直接反应用户的偏好（评分、喜欢/不喜欢）。但是实际场景中，用户对物品的评分数据是很难获取的，一般只有利用好隐式反馈数据，比如用户的各种行为数据（如点击、加购、收藏等等）。数据的类型不同，我们采取的处理方法和模型都会不一样。

# 大数据技术基础

## HDFS分布式存储

### Hadoop大数据平台

Hadoop是一个框架，它是由Java语言来实现的。Hadoop是处理大数据技术.  Hadoop可以处理云计算产生大数据，需要区分hadoop并不是云计算。它和云计算密不可分。详细见下面内容。

（1）Hadoop是如何产生的

Hadoop产生是互联网的产物，也是必然。大家都知道，我们上网时需要服务器的。假如世界上只有一台电脑，根本不需要服务器。如果有10台服务器，100台，1000台，上万台，那么我们该如何让大家相互通信，共享知识，所以我们产生了互联网。

互联网产生，全世界都可以通信，知识如此居多，我们像获取更多的知识，想获取新技术，获取新知识，通过什么，国内通过百度，国外也有许多，比如Google。可是百度和谷歌的用户有多少，多了不说，最起码有上亿的用户。并且这些用户每天上百度，上谷歌，又会产生多少数据，查询多少数据。那么他们怎么承受如此多用户。这不是一台电脑、一台服务器能完成的事情。

Hadoop就是一个解决方案。Hadoop是一个分布式方案，能够把压力分摊到其他服务器。至于如何做到的，可以深入了解Hadoop的maprecude等知识。想学习hadoop：可以查看下面内容：

云计算的部署方式

从部署方式来说，总共有两类云计算：

* 私有云：数据中心部署在企业内部，由企业自行管理。微软为大家提供了Dynamic Data Center Toolkit，来方便大家管理自己的数据中心。
* 公共云：数据中心由第三方的云计算供应商提供，供应商帮助企业管理基础设施（例如硬件，网络，等等）。企业将自己的软件及服务部属在供应商提供的数据中心，并且支付一定的租金。Windows Azure正是这样一个公共云平台。

具体来说，hadoop两个大的功能：海量数据的存储；海量数据的分析；

Hadoop有3大核心组件：

HDFS

---- hadoop分布式文件系统海量数据的存储(集群服务)，

MapReduce

----运算框架（导jar包写程序），海量数据运算分析（替代品：torm /spark/tez等 ）

Yarn

----资源调度管理集群(可以理解为一个分布式的操作系统，集群服务)

HDFS的架构与设计原理

分而治之—将大文件、大批量文件，分布式存放在大量独立的服务器上，以便于采取分而治之的方式对海量数据进行运算分析；

重点概念：文件切块，副本存放，元数据，位置查询，数据读写流

首先，它是一个文件系统，有一个统一的命名空间—目录树

其次，它是分布式的，由很多服务器联合起来实现功能；

1. hdfs文件系统会给客户端提供一个统一的抽象目录树，客户端访问hdfs文件时就是通过指定这个抽象目录中的路径来访问
2. Hdfs中的文件都是分块（block）存储的，块的大小可以通过配置参数( dfs.blocksize)来规定，默认大小在hadoop2.x版本中是128M，老版本中是64M
3. 文件的各个block由谁来进行真实的存储呢？----分布在各个datanode服务节点上，而且每一个block都可以存储多个副本（副本数量也可以通过参数设置dfs.replication）
4. Hdfs中有一个重要的角色：namenode，负责维护整个hdfs文件系统的目录树，以及每一个路径（文件）所对应的block块信息（block的id，及所在的datanode服务器）
5. hdfs是设计成适应一次写入，多次读出的场景，并不支持文件的修改

(hdfs并不适合用来做网盘应用，因为，不便修改，延迟大，网络开销大，成本太高)

### HDFS集群方式安装

（学习要求：掌握开发测试级别的hadoop集群部署运维）

1. 准备linux服务器（centos 6.4 32位<生产环境的服务器应该采用64位，可以支持更大的内存>）
2. 解压centos镜像压缩包到某个目录，并用vmware打开
3. 准备操作系统环境（主机名，ip地址配成static，域名和ip的本地映射hosts）

关闭图形界面的启动 修改/etc/inittab中的启动级别为3

1. 准备java环境，安装jdk，配置环境变量等

* 解压安装包，修改环境变量： JAVA\_HOME PATH
* 配置防火墙（关闭）
* 为hadoop软件准备一个专门的linux用户（在我们的linux系统镜像中，有用户hadoop，密码：hadoop），为hadoop用户设置sudo权限 /etc/sudoers

1. 安装hadoop----（解压，修改配置文件，分发到集群，初始化，启动）

Hadoop的目录结构:

|  |
| --- |
| bin #可执行文件（hadoop的功能操作命令）  etc #配置文件  include  lib #本地库文件（数据压缩编解码、本地文件系统操作）  libexec  LICENSE.txt  NOTICE.txt  README.txt  sbin #可执行文件（hadoop集群进程管理的操作命令）  share #开发所需要的jar包及用户帮助文档 |

1. 修改配置文件（参考现成的配置文件xxx-site.xml）

(1)hadoop-env.sh JAVA\_HOME = /home/hadoop/app/jdk\_7u65

(2)core-site.xml

fs.defaultFS 指定hadoop所使用的文件系统

hadoop.tmp.dir 指定各节点上的hadoop进程所在的本地工作目录(父目录)

(3) mapred-site.xml mapreduce.framework.name : yarn

(4)yarn-site.xml yarn.resourcemanager.hostname：server01 (yarn中的master节点所在主机)

yarn.nodemanager.aux-services : mapreduce\_shuffle

(5)可选：

如果要让namenode单独配置一个工作目录，在hdfs-site.xml ：

<name>dfs.namenode.name.dir</name>

<value> /mnt/driver-meta/,nfs://</value>

如果要让datanode单独配置一个工作目录，在hdfs-site.xml ：

<name>dfs.datanode.data.dir</name>

<value> /mnt/driver-data-a/,/mnt/driver-data-b/,/mnt/driver-data-c/</value>

如果要让secondary namenode在指定的机器上启动，则配置：

<name>dfs.namenode.secondary.http-address</name>

<value>hadoop-server02:50090</value>

(6)真实生产中部署一个中型集群：

有些公司会借助一些自动化的网络拷贝工具加快配置速度

有些公司会采用一些商业发行版（CDH—cloudera公司的产品；HORTONWORKS；MICROSOFT，IBM，EMC，INTEL）

1. 启动hadoop

首先，格式化nameonde bin/hadoop namenode -format

* 手动一台一台地启动

在相应服务器上启动hdfs的相关进程 ：

启动namenode进程—sbin/hadoop-daemon.sh start namenode

启动datanode进程—sbin/hadoop-daemon.sh start datanode

然后，验证hdfs的服务是否能正常提供：

bin/hdfs dfsadmin -report 查看hdfs集群的统计信息

* Shell脚本批量启动方式：

在任意一台服务器上执行命令：

启动hdfs服务：sbin/start-dfs.sh

启动yarn服务：sbin/start-yarn.sh

或者：直接启动hdfs+yarn服务： sbin/start-all.sh

6、集群内部的SSH密钥认证登陆机制配置(免密登陆)

配置的机制：在登陆方生成密钥对，然后将公钥复制给目标主机，在目标主机上将这个公钥加入授权文件 ~/.ssh/authorized\_keys (该文件的权限: 600)

真实大量配置的时候直接使用ssh工具箱的工具：

1/在登陆方生成密钥对，执行命令： ssh-keygen

2/执行这条指令：

|  |
| --- |
| ssh-copy-id hadoop-server03 |

就可以免密登陆目标主机

### HDFS的shell操作

* appendToFile ----追加一个文件到已经存在的文件末尾

Eg: hadoop fs -appendToFile ./hello.txt hdfs://hadoop-server01:9000/hello.txt

可以简写为：

Hadoop fs -appendToFile ./hello.txt /hello.txt

* cat ---显示文件内容

Eg: hadoop fs -cat /hello.txt

* chgrp
* chmod
* chown上面三个跟linux中的用法一样 Eg: hadoop fs -chmod 666 /hello.txt
* copyFromLocal #从本地文件系统中拷贝文件到hdfs路径去

Eg: hadoop fs -copyFromLocal ./jdk.tar.gz /aaa/

* copyToLocal #从hdfs拷贝到本地

Eg： hadoop fs -copyToLocal /aaa/jdk.tar.gz

* count #统计一个指定目录下的文件节点数量

Eg: hadoop fs -count /aaa/

* cp #从hdfs的一个路径拷贝hdfs的另一个路径 hadoop fs -cp /aaa/jdk.tar.gz /bbb/jdk.tar.gz.2
* createSnapshot
* deleteSnapshot
* renameSnapshot以上三个用来操作hdfs文件系统目录信息快照 Eg: hadoop fs -createSnapshot /
* df #统计文件系统的可用空间信息
* du Eg: hadoop fs -df -h / Eg: hadoop fs -du -s -h /aaa/\*
* get #等同于copyToLocal，就是从hdfs下载文件到本地
* getmerge #合并下载多个文件

Eg: 比如hdfs的目录 /aaa/下有多个文件:log.1, log.2,log.3,...

hadoop fs -getmerge /aaa/log.\* ./log.sum

* help #输出这个命令参数手册
* ls #显示目录信息

Eg: hadoop fs -ls hdfs://hadoop-server01:9000/

这些参数中，所有的hdfs路径都可以简写

Eg: hadoop fs -ls / 等同于上一条命令的效果

* mkdir #在hdfs上创建目录

Eg: hadoop fs -mkdir -p /aaa/bbb/cc/dd

* moveFromLocal #从本地剪切粘贴到hdfs
* moveToLocal #从hdfs剪切粘贴到本地
* mv #在hdfs目录中移动文件
* put #等同于copyFromLocal
* rm #删除文件或文件夹

Eg: hadoop fs -rm -r /aaa/bbb/

* rmdir #删除空目录
* setrep #设置hdfs中文件的副本数量

Eg: hadoop fs -setrep 3 /aaa/jdk.tar.gz

* stat #显示一个文件或文件夹的元信息
* tail #显示一个文件的末尾
* text #以字符形式打印一个文件的内容

### HDFS的java操作

可以随机定位读取位置：DFSInputStream.seek()

1. 搭建开发环境（eclipse，hdfs的jar包----hadoop的安装目录的share下）

|  |  |
| --- | --- |
| |  | | --- | | <dependency>  <groupId>org.apache.hadoop</groupId>  <artifactId>hadoop-client</artifactId>  <version>2.4.1</version>  </dependency> | |

建议在linux下进行客户端应用的开发，不会存在兼容性问题。

如果非要在window上做客户端应用开发，需要设置以下环境：

1. 在windows的某个目录下解压一个hadoop的安装包
2. 将安装包下的lib和bin目录用对应windows版本平台编译的本地库替换
3. 在window系统中配置HADOOP\_HOME指向你解压的安装包
4. 在windows系统的path变量中加入hadoop的bin目录
5. 在java中操作hdfs，首先要获得一个客户端实例

|  |
| --- |
| Configuration conf = new Configuration()  FileSystem fs = FileSystem.get(conf) |

而我们的操作目标是HDFS，所以获取到的fs对象应该是DistributedFileSystem的实例；

get方法是从何处判断具体实例化那种客户端类呢？

----从conf中的一个参数fs.defaultFS的配置值判断；

如果我们的代码中没有指定并且工程classpath下也没有给定相应的配置，conf中的默认值就来自于hadoop的jar包中的core-default.xml，默认值为： file:///

fs所具备的方法：

图片包含 屏幕截图

描述已自动生成

## Spark分布式计算平台

### Spark安装

Spark是一个快速而通用的集群计算平台。

在速度方面，Spark扩展了广泛使用的MapReduce计算模型，而且高效地支持更多计算模 式，包括交互式查询和流处理。在处理大规模数据集时，速度是非常重要的。速度快就意 味着我们可以进行交互式的数据操作，否则我们每次操作就需要等待数分钟甚至数小时。 Spark的一个主要特点就是能够在内存中进行计算，因而更快。不过即使是必须在磁盘上 进行的复杂计算，Spark依然比MapReduce更加高效。

总的来说，Spark适用于各种各样原先需要多种不同的分布式平台的场景，包括批处理、 迭代算法、交互式查询、流处理。通过在一个统一的框架下支持这些不同的计算，Spark使我们可以简单而低耗地把各种处理流程整合在一起。而这样的组合，在实际的数据分析 过程中是很有意义的。不仅如此，Spark的这种特性还大大减轻了原先需要对各种平台分 别管理的负担。

Spark所提供的接口非常丰富。除了提供基于Python、Java、Scala和SQL的简单易用的API以及内建的丰富的程序库以外，Spark还能和其他大数据工具密切配合使用。例如， Spark可以运行在Hadoop集群上，访问包括Cassandra在内的任意Hadoop数据源。

spark历史版本

spark存储层次

Spark不仅可以将任何Hadoop分布式文件系统(HDFS)上的文件读取为分布式数据集， 也可以支持其他支持Hadoop接口的系统，比如本地文件、亚马逊S3、Cassandra、Hive、 HBase等。我们需要弄清楚的是，Hadoop并非Spark的必要条件，Spark支持任何实现 了Hadoop接口的存储系统。Spark支持的Hadoop输入格式包括文本文件、SequenceFile、 Avro、Parquet等。我们会在第5章讨论读取和存储时详细介绍如何与这些数据源进行交互。

=====> 可以说一下spark的快读数据存储格式，parquet等等！！！

spark下载与安装

使用Spark的第一步是下载和解压缩。我们先从下载预编译版本的Spark开始。访问http:// spark.apache.org/downloads.html， 选 择 包 类 型 为“Pre-built for Hadoop 2.4 and later”( 为Hadoop 2.4及更新版本预编译的版本)，然后选择“Direct Download”直接下载。这样我们 就可以得到一个压缩的TAR文件，文件名为spark-1.2.0-bin-hadoop2.4.tgz.

你不需要安装Hadoop，不过如果你已经有了一个Hadoop集群或安装好的HDFS，请下 载对应版本的Spark。你可以在<http://spark.apache.org/downloads.html>里选择所需要的包 类型，这会导致下载得到的文件名略有不同。也可以选择从源代码直接编译。你可以从GitHub上下载最新代码，也可以在下载页面上选择包类型为“Source Code”(源代码)进 行下载。

下载好了Spark之后，我们要进行解压缩，然后看一看默认的Spark发行版中都有些什么。 打开终端，将工作路径转到下载的Spark压缩包所在的目录，然后解开压缩包。这样会创 建出一个和压缩包同名但是没了 .tgz后缀的新文件夹。接下来我们就把工作路径转到这个 新目录下看看里面都有些什么。上面这些步骤可以用如下命令完成:

cd ~ tar -xf spark-1.2.0-bin-hadoop2.4.tgz

cd spark-2.4.0-bin-hadoop2.7

ls

在tar命令所在的那一行中，x标记指定tar命令执行解压缩操作，f标记则指定压缩包的文件名。ls命令列出了Spark目录中的内容。我们先来粗略地看一看Spark目录中的一些 比较重要的文件及目录的名字和作用。

Spark-shell

Spark带有交互式的shell，可以作即时数据分析。如果你使用过类似R、Python、Scala所 提供的shell，或操作系统的shell(例如Bash或者Windows中的命令提示符)，你也会对Spark shell感到很熟悉。然而和其他shell工具不一样的是，在其他shell工具中你只能使 用单机的硬盘和内存来操作数据，而Spark shell可用来与分布式存储在许多机器的内存或 者硬盘上的数据进行交互，并且处理过程的分发由Spark自动控制完成。

由于Spark能够在工作节点上把数据读取到内存中，所以许多分布式计算都可以在几秒钟 之内完成，哪怕是那种在十几个节点上处理TB级别的数据的计算。这就使得一般需要在shell中完成的那些交互式的即时探索性分析变得非常适合Spark。Spark提供Python以及Scala的增强版shell，支持与集群的连接。

### Spark编程基础

#### 开发环境搭建

RDD编程

Spark构建Maven项目

#### 基础数据类型

**RDD**

弹性分布式数据集（RDD），Spark中的基本抽象。表示为可以并行操作的数据集合，这个集合实际上是不可变的，并且以分区（Partition）的方式存在。此类包含所有RDD上可用的基本操作，例如地图，过滤器和持久性。

另外，PairRDDFunctions包含仅在键-值对的RDD上可用的操作，例如groupByKey和join； DoubleRDDFunctions包含仅在Doubles的RDD上可用的操作； SequenceFileRDDFunctions包含RDD上可用的操作，这些操作可以另存为SequenceFiles。通过隐式转换，这些操作可自动用于任何正确类型的RDD（例如RDD[(Int, Int)]）。

从功能上将，SparkContext是Spark程序主入口，而RDD代表了分布式集合的数据类型，并且提供各种并行操作。

进入spark安装目录，打开spark-shell

./bin.spark-shell

读取本地文件README.md

val data = spark.sparkContext.textFile(“README.md”)

data: org.apache.spark.rdd.RDD[String] = README.md MapPartitionsRDD[1] at textFile at <console>:23

查看数据的第一行：  
data.first()

Res0: String = # Apache Spark

统计数据的行数：  
data.count()

res1: Long = 105

**Dataset**

DataSet是分布式的数据集合。DataSet是在Spark1.6之后新增加的接口。它集中了RDD的优点（强类型输入，并且支持强大lambda函数）以及Spark SQL优化的执行引擎。DataSet可以通过JVM的对象进行构建，可以用函数式的转换（map/flatmap/filter等等）进行多种操作。

Dataset是“惰性”的，即仅在调用动作时才触发计算。在内部，数据集表示描述生成数据所需的计算的逻辑计划。调用操作时，Spark的查询优化器将优化逻辑计划并生成物理计划，以并行和分布式方式高效执行。要探索逻辑计划以及优化的物理计划，请使用解释功能。 为了有效地支持特定于域的对象，需要一个编码器Encoder。编码器将特定于域的类型T映射到Spark的内部类型系统。例如，给定Person类具有name(String) 和age(Int) 两个l字段，则使用编码器告诉Spark在运行时生成代码以将Person对象序列化为二进制结构。该二进制结构通常具有低得多的存储器占用空间，并且针对数据处理的效率进行了优化（例如，以列格式）。要了解数据的内部二进制表示形式，请使用架构函数。 通常有两种创建数据集的方法。最常见的方法是使用SparkSession上可用的读取功能，将Spark指向存储系统上的某些文件。

scala代码读取parquet文件如下：

case class Person(name: String, age: Int)

val people = spark.read.parquet(“...”).as[Person]

也可以通过现有Dataset上适用的转换操作来创建Dataset，比如通过在现有过滤器上来创建新的Dataset。

//对Dataset[People]进行map操作的scala代码

val names = people.map(\_.name)

Dataset操作可以定义为无类型Dataset、列、函数，这种操作与Python中的dataframe类型十分相似。

下面scala代码中直接调用方式选择一列数据：

val ageCol = people(“age”)

Dataset中列可以通过各种方法进行操作。比如下面在age列的基础上加上10，并创造一个新的列。 people(“age”) + 10 // in Scala

在看一个更加复杂的例子。使用SparkSession创建一个Dataset[Row]。使用department作为另一个数据源，它的id字段与people中的deptId字段关联。

val people = spark.read.parquet(“...”)  
val department = spark.read.parquet(“...”)   
people.filter(“age > 30”).  
 join(department, people(“deptId”) === department(“id”)).  
 groupBy(department(“name”), people(“gender”)).  
 agg(avg(people(“salary”)), max(people(“age”)))

**DataFrame**

DataFrame和DataSet类似，也是个分布式集合，其中数据别组织成命名的列，可以看做关系数据库中的表，底层做了很多优化，可以通过很多数据源进行构建，比如RDD、结构化文件、外部数据库、Hive表。

DataFrame的前身是SchemaRDD。Spark1.3开始SchemaRDD更改为DataFrame。区别，不继承RDD，自己实现了RDD的大部分功能。可以在DataFrame上调用RDD的方法转化成另外一个RDD。

DataFrame可以看做分布式Row对象的集合，其提供了由列组成的详细模式信息，

使其可以得到优化。

DataFrame是通过列名组成的Dataset。从概念上讲，它等效于关系数据库中的表或R / Python中的数据框，但是在后台进行了更丰富的优化。可以从多种来源构造DataFrame，例如：结构化数据文件，Hive中的表，外部数据库或现有RDD。 DataFrame API在Scala，Java，Python和R中可用。在Scala和Java中，DataFrame由行的数据集表示。在Scala API中，DataFrame只是Dataset [Row]的类型别名。

总结：

DataFrame不仅有比RDD更多的算子，还可以进行执行计划的优化。

DataSet包含了DataFrame的功能，Spark2.0中两者统一，DataFrame表示为DataSet[Row]，即DataSet的子集。所以，使用API尽量使用DataSet[[2]](#footnote-2)。

#### Transformer算子

#### Action算子

### Spark MLlib数据类型

Spark MLlib底层的向量和矩阵运算使用了Breeze库，Breeze库提提供了Vector和Matrix的实现以及相关的计算接口。但是在MLlib里面同时也提供了Vector和Linalg等的实现。在MLlib函数里的参数传递均使用MLlib自己的Vector，而且在函数内的矩阵计算又通过ToBreeze.ToDenseVector变成Breeze的形式进行计算。这样做的目的一是保持函数接口的稳定性，不会因为Breeze的变化而变化；另外一个比较重要的就是可以把Distributed Matrix作为一种Matrix的实现而被使用。

Spark在Vector和Matrix的基础上，实现了分布式矩阵类。分布式矩阵的数据分块或者分行存储，并且实现了矩阵的基本运算，能够使矩阵分布式计算，比如列统计、相似度、协方差、奇异值分解等。

#### 本地向量（Local Vector）

本地向量是从0开始的下标和double类型的值，存储本地机器中，所以称为Local Vector。其支持两种数据形式：

Dense（稠密的向量）

Sparse（稀疏的向量）

比如一个向量[1.0,0.0,3.0]，用Dense表示为：[1.0,0.0,3.0]，用Sparse表示为：(3,[0,2],[1.0,3.0])，其中3为向量的长度，[0,2]表示元素[1.0,3.0]的位置，可见sparse形式下0.0是不存储的。

import org.apache.spark.mllib.linalg.{Vector, Vectors}  
  
// 创建稠密向量 (1.0, 0.0, 3.0).  
val dv: Vector = Vectors.dense(1.0, 0.0, 3.0)

//指定下标和数值，创建稀疏向量 (1.0, 0.0, 3.0)   
val sv1: Vector = Vectors.sparse(3, Array(0, 2), Array(1.0, 3.0))

//指定所有非零元素来创建稀疏向量 (1.0, 0.0, 3.0)，

val sv2: Vector = Vectors.sparse(3, Seq((0, 1.0), (2, 3.0)))

#### 标记点向量（Labeled Point）

labeled point由本地向量组成，既可以是dense向量，也可以是sparse向量。在mllib中常用于监督类算法，使用double类型来保存该类型的数据，因为也可以用于回归和分类算法。例如二分类，label可以是0（负例）或1（正例），对于多分类，label可以是0，1，2…

import org.apache.spark.mllib.linalg.Vectors

import org.apache.spark.mllib.regression.LabeledPoint

//用一个正标签和一个稠密特征向量创建标记点

val pos = LabeledPoint(1.0, Vectors.dense(1.0, 0.0, 3.0))

// Create a labeled point with a negative label and a sparse feature vector.

//用一个负标签和一个稀疏特征向量创建标记点

val neg = LabeledPoint(0.0, Vectors.sparse(3, Array(0, 2), Array(1.0, 3.0)))

#### 本地矩阵（Local Matrix）

本地矩阵由行下标，列索引和double类型的值组成，存储在本地机器上，mllib支持稠密矩阵和稀疏矩阵，其存储是按照列进行存储的。

比如有一个稠密矩阵： 通过数组存储的形式为： [1.0, 3.0, 5.0, 2.0, 4.0, 6.0]，矩阵大小为[3，2]。

// 创建稠密矩阵 ((1.0, 2.0), (3.0, 4.0), (5.0, 6.0))

Import org.apache.spark.mllib.linalg.Matrices

val denseMatrix = Matrices.dense(3,2, Array(1.0,3.0,5.0,2.0,4.0,6.0))

println(s”denseMatrix is : $denseMatrix”)

输出结果如下：

denseMatrix is :

1.0 2.0

3.0 4.0

5.0 6.0

// 创建稀疏矩阵 ((1.0, 0.0, 4.0), (0.0, 3.0, 5.0), (2.0, 0.0, 6.0))

val sparseMatrix = Matrices.sparse(3,3, Array(0,2,3,6),Array(0,2,1,0,1,2),Array(1,2,3,4,5,6))

println(s”sparseMatrix is : $sparseMatrix”)

输出结果如下：

sparseMatrix is : 3 x 3 CSCMatrix

(0,0) 1.0

(2,0) 2.0

(1,1) 3.0

(0,2) 4.0

(1,2) 5.0

(2,2) 6.0

再比如一个稀疏矩阵 首先指定矩阵是3行3列，

数组: [1.0, 2.0, 3.0, 4.0, 5.0, 6.0]，

行下标（Row Indices）: [0, 2, 1, 0, 1, 2]，

列偏移（column offsets）：[0, 2, 3, 6]

Array(0, 2, 1, 0, 1, 2)是指，第一个非零元素在第0行，第二个非零元素在第2行，第三个非零元素在第1行，以此类推。

Array(0, 2, 3, 6)标记列偏移和元素数量，列偏移0指的的是第一列的1.0从0偏移量开始，偏移1指第二列3.0从数组偏移2开始，3指4.0开始的偏移量。6是非零元素数量。

此处设计比较好，假设10000个元素分两列，不需要把每个元素所在列都标出来，只需要记录6个数字即可。Array(1, 2, 3, 4, 5)表示按顺序存储非零元素。

### Spark分布式矩阵

一个分布式矩阵由下标和double类型的数据组成，不过分布式的矩阵的下标不是int类型，而是long类型，数据保存在一个或多个rdd中，选择一个正确的格式去存储分布式矩阵是非常重要的。分布式矩阵转换成不同的格式需要一个全局的shuffle(global shuffle)，而全局shuffle的代价会非常高。到目前为止，Spark MLlib中已经实现了三种分布式矩阵。

最基本的分布式矩阵是RowMatrix，它是一个行式的分布式矩阵，没有行索引。比如一系列特征向量的集合。RowMatrix由一个RDD代表所有的行，每一行是一个本地向量。假设一个RowMatrix的列数不是特别巨大，那么一个简单的本地向量能够与driver进行联系，并且数据可以在单个节点上保存或使用。IndexedRowMatrix与RowMatrix类似但是有行索引，行索引可以用来区分行并且进行连接等操作。CoordinateMatrix是一个以协同列表（coordinate list）格式存储数据的分布式矩阵，数据以RDD形式存储。

Spark实现了分布式矩阵类DistributedMatrix，它是Spark中所有分布式矩阵的父类。目前实现了一些分布式矩阵子类包括： RowMatrix，IndexedRowMatrix，CoordinateMatrix和BlockMatrix。

#### 行矩阵（RowMatrix）

分布式行矩阵就是把每行对应为一个RDD，将矩阵的每行分布式存储，矩阵的每行是一个本地向量。这和多变量统计的数据矩阵比较相似度。因为每行以一个本地向量表示，所以矩阵列的数量被限制在整数范围内，但是实际应用中的列数很小。

一个RowMatrix可以由一个RDD[Vector]的实例创建。因此我们可以计算统计信息或者进行分解。QR分解（QR decomposition）是A=QR，其中Q是一个矩阵，R是一个上三角矩阵。对sigular value decomposition(SVD和principal component analysis（PCA）,可以去参考降维的部分。

import org.apache.spark.mllib.linalg.Vectors

import org.apache.spark.mllib.linalg.distributed.RowMatrix

val arr = Array(Vectors.dense(1,0),Vectors.dense(0,1))

val rows = spark.sparkContext.parallelize(arr)

val mat: RowMatrix = new RowMatrix(rows)

val m = mat.numRows()

val n = mat.numCols()

val qrResult = mat.tallSkinnyQR(true)

println(s”m is: $m，n is $n，\nqrResult is :”)

qrResult.Q.rows.foreach(println)

println()

qrResult.R.rowIter.foreach(println)

其输出为：

Row Matrix ...

m is: 2，n is 2，

qrResult is :

[1.0,0.0]

[0.0,1.0]

[1.0,0.0]

[0.0,1.0]

#### 行索引矩阵（IndexedRowMatrix）

IndexRowMatrix和RowMatrix非常相似，区别是它带有一定意义的行索引。在RowMatrix中，rows的格式是RDD[Vector]；而在IndexedRowMatrix中，rows的格式是RDD[IndexRow]，其中的IndexedRow的格式是（index:Long, vector:Vector），相比RowMatrix多了一个Index索引信息。

一个IndexedMatrix可以从RDD[IndexedRow]实例创建，IndexedRow是（Long,Vector）的wrapper，而且这种矩阵可以转换成RowMatrix（通过去掉Index），其创建及使用方法类似于RowMatrix。

// IndexedRowMatrix

println(“Indexed Row Matrix ...”)

val arr2 = Array(

IndexedRow(0,Vectors.dense(1,0)),

IndexedRow(1,Vectors.dense(0,1))

)

val rows2: RDD[IndexedRow] = spark.sparkContext.parallelize(arr2)

val mat2 = new IndexedRowMatrix(rows2)

val m2 = mat2.numRows()

val n2 = mat2.numCols()

// 去掉行索引，转换成RowMatrix

val qrResult2 = mat2.toRowMatrix().tallSkinnyQR(true)

println(s”m2 is: $m2，n2 is $n2，\nqrResult2 is :”)

qrResult2.Q.rows.foreach(println)

println()

qrResult2.R.rowIter.foreach(println)

输出为：

Indexed Row Matrix ...

m2 is: 2，n2 is 2，

qrResult2 is :

[1.0,0.0]

[0.0,1.0]

[1.0,0.0]

[0.0,1.0]

#### 坐标矩阵（CoordinateMatrix）

坐标矩阵，每一项都是一个（i:Long, j:Long, value:Double）指示行列值的元组tuple。其中i是行坐标，j是列坐标，value是值。如果矩阵非常大并且稀疏，那么坐标矩阵一定是最好的选择。坐标矩阵是通过RDD[MatrixEntry]实例创建的，MatrixEntry为（Long, Long, Double）的形式。坐标矩阵可以转换为IndexedRowMatrix。

columnsSimilarities(threshold: Double)

threshold: 设置为0表示确定性的保证正确性。如果>0，结果相似度与上述成本与估计质量权衡相关。

返回一个n\*n稀疏的上三角矩阵中列与列之间的的余弦相似度。

// CoordinateMatrix

println(“Coordinate Matrix ...”)

val arr3 = Array(

MatrixEntry(0,0,1),

MatrixEntry(1,1,1)

)

val entries = spark.sparkContext.parallelize(arr3)

val mat3 = new CoordinateMatrix(entries)

val m3 = mat.numRows()

val n3 = mat.numCols()

val qrResult3 = mat3.toIndexedRowMatrix().toRowMatrix().tallSkinnyQR(true)

println(s”m3 is: $m3，n3 is $n3，\nqrResult3 is :”)

qrResult3.Q.rows.foreach(println)

println()

qrResult3.R.rowIter.foreach(println)

输出为：

Coordinate Matrix ...

m3 is: 2，n3 is 2，

rowMat3 is :

[1.0,0.0]

[0.0,1.0]

[1.0,0.0]

[0.0,1.0]

#### 分块矩阵（BlockMatrix）

分块矩阵（分段矩阵）就是将矩阵分割成较小的矩形矩阵，这些较小的矩阵就是区块。也就是说，以较小的矩阵组合成一个矩阵。分块矩阵的分割原则是以水平线和垂直线进行区分。在分块矩阵中，位于同一行（列）的每个子矩阵，都拥有相同的列数（行数）。将大的矩阵通过分块的方式划分，每个分块看做另一个矩阵的元素，再参与运算，可以让计算变得大幅简化。例如，有的大矩阵可以通过分块变成对角矩阵或者是三角矩阵等特殊形式的矩阵。

一个BlockMatrix是一个分布式的矩阵，由一个MatrixBlocks的RDD组成。MatrixBlock是一个三元组((Int, Int), Matrix),其中(Int, Int)是block的索引，Matrix是一个在指定位置上的维度为rowsPerBlock \* colsPerBlock的子矩阵。BlockMatrix支持与另一个BlockMatrix对象的add和multiply操作。BlockMatrix提供了一个帮助方法validate，这个方法可以用于检测该‘BlockMatrix·是否正确。

如矩阵：可以分成的区块，如下：

分块后的矩阵可以写作:

分块矩阵BlockMatrix的存储是以单位块blocks分布式存储的，blocks的格式为RDD[((blockRowIndex,blockColIndex),sub-matrix)]。其中，blockRowIndex是block块行索引，blockColIndex是block块列索引，sub-matrix是block块矩阵。

可以通过IndexedRowMatrix或者CoordinateMatrix调用toBlockMatrix快速得到BlockMatrix对象。默认情况下toBlockMatrix方法会得到一个1024 x 1024的BlockMatrix。使用时可以通过手动传递维度值来设置维度，toBlockMatrix(rowsPerBlock, colsPerBlock)。

分块矩阵的Scala代码案例：

// BlockMatrix

println(“Block Matrix ...”)

val arr4 = Array(

MatrixEntry(0,0,1),

MatrixEntry(1,1,1)

)

val entries4: RDD[MatrixEntry] = spark.sparkContext.parallelize(arr4)

val coordMat: CoordinateMatrix = new CoordinateMatrix(entries4)

val matA: BlockMatrix = coordMat.toBlockMatrix().cache()

// 检测BlockMatrix格式是否正确，错误的话会抛出异常，正确的话无其他影响

matA.validate()

matA.blocks.foreach(println)

val m4 = matA.numRowBlocks

val n4 = matA.numColBlocks

println(s”m4 is: $m4，n4 is $n4”)

// 计算A^T \* A.

val ata = matA.transpose.multiply(matA)

ata.blocks.foreach(println)

输出为：

Block Matrix ...

((0,0),2 x 2 CSCMatrix

(0,0) 1.0

(1,1) 1.0)

m4 is: 1，n4 is 1

((0,0),1.0 0.0

0.0 1.0 )

### 使用Spark进行数据预处理

为了让原始数据可用于机器学习算法，我们需要先对其进行清理，并且可能需要进行各种转 换，之后才能从转换后的数据里提取有用的特征。数据的转换和特征提取联系紧密。某些情况下，一些转换本身便是特征提取的过程。

在之前处理电影数据集时，我们已经看到数据清理的必要性。一般来说，现实中的数据会存 在信息不规整、数据点缺失和异常值问题。理想情况下，我们会修复非规整数据。但很多数据集都源于一些难以重现的收集过程(比如网络活动数据和传感器数据)，故实际上会难以修复。值缺失和异常也很常见，且处理方式可与处理非规整信息类似。总的来说，大致的处理方法如下。

过滤掉或删除非规整或有值缺失的数据:这通常是必需的，但的确会损失这些数据中那 些好的信息。 **10**

填充非规整或缺失的数据:可以根据其他的数据来填充非规整或缺失的数据。方法包括 用零值、全局平均值或中值来填充，或是根据相邻或类似的数据点来做插值(通常针对 时序数据)等。选择正确的方式并不容易，它会因数据、应用场景和个人经验而不同。

对异常值做稳健处理:异常值的主要问题在于，即使它们是极值也不一定就是错的。到 底是对是错通常很难分辨。异常值可被移除或填充，但存在某些统计技术(如稳健回归) 可用于处理异常值或是极值。

对可能的异常值进行转换:另一种处理异常值或极值的方法是进行转换。对那些可能存 在异常值或值域覆盖过大的特征，进行对数或高斯核转换。这类转换有助于降低变量存 在的值跳跃的影响，并将非线性关系变为线性的。

异常数据与缺失数据的填充

### Spark MLlib特征处理

初步完成源数据的清洗、处理和探索之后，就可以尝试从数据中提取供模型训练使用的特征。特征主要是指作为模型训练的变量，特征在模型训练之前需要转换成若干行的的向量，向量的元素就是数值。

特征可以概括地分为如下几种。

数值特征(numerical feature):这些特征通常为实数或整数，比如之前例子中提到的年龄。

类别特征(categorical feature):它们的取值只能是可能状态集合中的某一种。我们数据集中的用户性别、职业或电影类别便是这类特征。

文本特征(text feature):它们派生自数据中的文本内容，比如电影名、描述或评论。

其他特征:大部分其他特征最终都表示为数值。比如图像、视频和音频可被表示为数值

数据的集合，地理位置则可由经纬度或地理散列(geohash)表示。 本节我们将谈到数值特征、类别特征以及文本特征。

#### 特征抽取

**数值特征**

原始的数值和数值特征之间的区别是什么?实际上，任何数值数据都能作为输入变量。但是，机器学习模型中所学习的是各个特征的权重向量。权重在特征值到输出或目标变量(指在监督学 习模型中)的映射过程中扮演重要角色。

由此我们会想使用那些合理的特征，让模型能从这些特征中学到特征值和目标变量之间的关 系。比如年龄就是一个合理的特征。年龄的增加和某项支出之间可能存在直接关系。类似地，高 度也是一个可直接使用的数值特征。

当数值特征仍处于原始形式时，其可用性相对较低，但可以转化为更有用的表示形式。位置信息便是如此。

若使用原始位置信息(比如用经纬度表示的)，我们的模型可能学习不到该信息和某个输出 之间的有用关系，这就使得该信息的可用性不高，除非数据点的确很密集。然而，对位置进行聚合或挑选后(比如聚焦为一个城市或国家)，便可能和特定输出之间存在某种关联。

**类别特征**

当类别特征仍为原始形式时，其取值来自所有可能取值所构成的集合，而不是一个数字， 故不能作为输入。之前例子中的用户职业便是一个类别特征变量，其可能的取值有学生、程序 员等。

这样的类别特征也称作名义(nominal)变量，即其各个可能取值之间没有顺序关系。相反， 那些存在顺序关系的(比如之前提到的评级，从定义上说评级5会高于或是好于评级1)则被称**5**为有序(ordinal)变量。

将类别特征表示为数字形式，常可借助*k*之1(1-of-*k*)编码方法进行。要想将名义变量表示 为可用于机器学习任务的形式，需要借助如*k*之1编码这样的方法。有序变量的原始值可能就能 直接使用，但也常会经过和名义变量一样的编码处理。

假设变量可取的值有*k*个。如果对这些值用1到*k*编序(索引)，则可以用长度为*k*的二元 向量来表示该变量的一个取值。在这个向量里，该取值对应的序号所在的元素为1，其他元素 都为0。

比如职业student对应的索引为0，programmer对应的索引为1。那么student和programmer对应的值就分别为[1,0]和[0,1]。

**派生特征**

前面曾提到，从现有的一个或多个变量派生出新的特征常常是有帮助的。理想情况下，派生出的特征能比原始变量带来更多信息。

比如，可以分别计算各用户已有的电影评级的平均数。这将能给模型加入针对不同用户的个 性化特征(事实上，这常用于推荐系统)。在前文中我们也从原始的评级数据创建了新的特征以 学习出更好的模型。

从原始数据派生特征的例子包括计算平均值、中位值、方差、和、差、最大值或最小值以及 计数。在先前的内容中，我们已看到如何从电影的发行年份和当前年份派生出了新的movie age特征。这类转换背后的想法常常是对数值数据进行某种概括，让模型更容易学习这些特征。

数值特征到类别特征的转换也很常见，比如划分为区间特征。进行这类转换的常见变量有年 龄、地理位置和时间。

**文本特征[[3]](#footnote-3)**

Spark MLlib提供三种文本特征提取方法，分别为TF-IDF、Word2Vec和CountVectorizer。

**计数向量器CountVectorizer**

计数向量器CountVectorizer和计数向量器模型CountVectorizerModel是通过计数的方式将文档转换成向量。当不存在先验词典时，CountVectorizer可以用作Estimator来提取词汇表并生成CountVectorizerModel。该模型为词汇表上的文档生成稀疏表示，可以将其传递给其他算法，例如LDA[[4]](#footnote-4)。

在拟合过程中，CountVectorizer将选择整个语料库中按词频排列的前vocabSize大小的词。参数minDF通过指定一个单词必须出现在词汇表中的最小数量（如果小于1.0，则为小数）来影响拟合过程。另一个可选的二进制切换参数控制输出向量。如果将其设置为true，则所有非零计数都将设置为1。这对于模拟二进制而不是整数计数的离散概率模型特别有用。

代码示例：

import org.apache.spark.ml.feature.{CountVectorizer, CountVectorizerModel}

val df = spark.createDataFrame(Seq(

(0, Array(“a”, “b”, “c”)),

(1, Array(“a”, “b”, “b”, “c”, “a”))

)).toDF(“id”, “words”)

// fit a CountVectorizerModel from the corpus

val cvModel: CountVectorizerModel = new CountVectorizer()

.setInputCol(“words”)

.setOutputCol(“features”)

.setVocabSize(3)

.setMinDF(2)

.fit(df)

// alternatively, define CountVectorizerModel with a-priori vocabulary

val cvm = new CountVectorizerModel(Array(“a”, “b”, “c”))

.setInputCol(“words”)

.setOutputCol(“features”)

cvModel.transform(df).select(“features”).show()

假设我们有如下的DataFrame包含id和texts两列：

id | texts

0 |Array(“a”, “b”, “c”)

1 |Array(“a”, “b”, “b”, “c”,”a”)

文本中的每一行都是一个文档类型的数组(字符串)。调用的CountVectorizer产生词汇(a,b,c)的CountVectorizerModel，转换后的输出向量如下：

id | texts | vector

0 |Array(“a”, “b”, “c”) | (3,[0,1,2],[1.0,1.0,1.0])

1 |Array(“a”, “b”, “b”, “c”,”a”) |(3,[0,1,2],[2.0,2.0,1.0])

每个向量代表文档的词汇表中每个词语出现的次数。

**TF-IDF**

TF-IDF是一种广泛用于文本挖掘中的特征向量化方法，它可以反映术语对语料库中文档的重要性。

词语由t表示，文档由d表示，语料库由D表示。

其中，表示词在文件中出现的次数，而分母表示在文件中所有词出现次数之和。

词频TF(t,d)是词语t在文档d中出现的次数。文件频率DF(t,D)是包含词语的文档的个数。如果我们只使用词频来衡量重要性，很容易过度强调在文档中经常出现而并没有包含太多与文档有关的信息的词语，比如“一个”，“该”和“属于”等等。如果术语经常出现在整个语料库中，则表示该术语不包含有关特定文档的特殊信息。逆文档频率(IDF)是一个术语多少信息提供了一个数值量度，IDF的数值化衡量词语提供多少信息，用以下公式表示：

其中是语料库中的文档的总数。由于使用对数，因此如果一个术语出现在所有文档中，则其IDF值将变为0。请注意，应用了平滑术语以避免对主体外的术语除以零。TF-IDF度量只是TF和IDF的乘积：

​在Spark ML库中，TF-IDF被分成两部分：TF (+hashing) 和IDF[[5]](#footnote-5)。

HashingTF是一个Transformer，在文本处理中，接收词条的集合然后把这些集合转化成固定长度的特征向量。这个算法在哈希的同时会统计各个词条的词频。

IDF是一个Estimator，在一个数据集上应用它的fit（）方法，产生一个IDFModel。 该IDFModel接收特征向量（由HashingTF产生），然后计算每一个词在文档中出现的频次。IDF会减少那些在语料库中出现频率较高的词的权重。

​Spark.mllib中实现词频率统计使用特征hash的方式，原始特征通过hash函数，映射到一个索引值。后面只需要统计这些索引值的频率，就可以知道对应词的频率。这种方式避免设计一个全局1对1的词到索引的映射，这个映射在映射大量语料库时需要花费更长的时间。但需要注意，通过hash的方式可能会映射到同一个值的情况，即不同的原始特征通过Hash映射后是同一个值。为了降低这种情况出现的概率，我们只能对特征向量升维。i.e., 提高hash表的桶数，默认特征维度是2^20 = 1,048,576.

在下面的代码段中，我们以一组句子开始。首先使用分解器Tokenizer把句子划分为单个词语。对每一个句子（词袋），我们使用HashingTF将句子转换为特征向量，最后使用IDF重新调整特征向量。这种转换通常可以提高使用文本特征的性能。然后，我们的特征向量可以在算法学习中。

import org.apache.spark.ml.feature.{HashingTF, IDF, Tokenizer}

val sentenceData = spark.createDataFrame(Seq(

(0, “Hi I heard about Spark”),

(0, “I wish Java could use case classes”),

(1, “Logistic regression models are neat”)

)).toDF(“label”, “sentence”)

val tokenizer = new Tokenizer().setInputCol(“sentence”).setOutputCol(“words”)

val wordsData = tokenizer.transform(sentenceData)

从打印结果可以看到，tokenizer的transform()方法把每个句子拆分成了一个个单词。

val hashingTF = new HashingTF()

.setInputCol(“words”).setOutputCol(“rawFeatures”).setNumFeatures(20)

val featurizedData = hashingTF.transform(wordsData)

我们可以看到每一个单词被哈希成了一个不同的索引值。以”I heard about Spark and I love Spark”为例，输出结果中2000代表哈希表的桶数，“[105,365,727,1469,1858,1926]”分别代表着“i, spark, heard, about, and, love”的哈希值，“[2.0,2.0,1.0,1.0,1.0,1.0]”为对应单词的出现次数。

// alternatively, CountVectorizer can also be used to get term frequency vectors

调用IDF方法来重新构造特征向量的规模，生成的idf是一个Estimator，在特征向量上应用它的fit（）方法，会产生一个IDFModel。

val idf = new IDF().setInputCol(“rawFeatures”).setOutputCol(“features”)

val idfModel = idf.fit(featurizedData)

val rescaledData = idfModel.transform(featurizedData)

rescaledData.select(“features”, “label”).take(3).foreach(println)

​“[105,365,727,1469,1858,1926]”分别代表着“i, spark, heard, about, and, love”的哈希值。105和365分别代表了“i”和”spark”，其TF-IDF值分别是0.2876820724517808和0.6931471805599453。这两个单词都在第一句中出现了两次，而”i”在第二句中还多出现了一次，从而导致”i”的TF-IDF度量值较低。因此，与“i”相比，“spark”能更好的区分文档。

**Word2Vec**

Word2Vec是一个Estimator表示代表文档的单词序列来训练一个Word2VecModel。该模型将每个单词映射到唯一的固定大小的向量。Word2VecModel使用文档中所有单词的平均值将每个文档转换为向量。之后，将此向量用作特征进行预测，或者用以计算文档相似度。

下面代码中，假设一组文档，其中每个文档代表一个词语序列。将每个文档转换成特征向量，然后将特征向量传递到学习算法中。

import org.apache.spark.ml.feature.Word2Vec

// Input data: Each row is a bag of words from a sentence or document.

val documentDF = spark.createDataFrame(Seq(

“Hi I heard about Spark”.split(“ “),

“I wish Java could use case classes”.split(“ “),

“Logistic regression models are neat”.split(“ “)

).map(Tuple1.apply)).toDF(“text”)

// Learn a mapping from words to Vectors.

val word2Vec = new Word2Vec()

.setInputCol(“text”)

.setOutputCol(“result”)

.setVectorSize(3)

.setMinCount(0)

val model = word2Vec.fit(documentDF)

val result = model.transform(documentDF)

result.select(“result”).take(3).foreach(println)

**正则化特征**

#### 特征转换

在机器学习中，数据处理是一件比较繁琐的事情，需要对原有特征做多种处理，如类型转换、标准化特征、新增衍生特征等等，需要耗费大量的时间和精力编写处理程序，不过，自从Spark推出ML后，情况大有改观，Spark ML包中提供了很多现成转换器，例如：StringIndexer、IndexToString、OneHotEncoder、VectorIndexer，它们提供了十分方便的特征转换功能，这些转换器类都位org.apache.spark.ml.feature包下。

**规范化Normalizer[[6]](#footnote-6)**

将某个特征向量（由所有样本某一个特征组成的向量）计算其p-范数，然后对该每个元素除以p-范数。将原始特征Normalizer以后可以使得机器学习算法有更好的表现。

单位P-范数定义如下：

**StandardScaler**

z−score规范化，又叫零均值规范化

将某个特征向量（由所有样本某一个特征组成的向量）进行标准化，使数据均值为0，方差为1。Spark中可以选择是带或者不带均值和方差。

注意：尤其是离群点左右了MinMaxScaler规范化,需要使用StandardScaler。

\_

Spark中有两个参数可以选择：

1、withStd=true,将方差缩放到1，

2、withMean-将均值移到0，注意对于稀疏输入矩阵不可以用。默认为false。

**MinMaxScaler**

最大-最小规范化：

将所有特征向量线性变换到用户指定最大-最小值之间。但注意在计算时还是一个一个特征向量分开计算的（见下面公式）通常将最大，最小值设置为1和0，这样就归一化到[0,1]。Spark中可以对min和max进行设置，默认就是[0,1]。

注意：（1）最大最小值可能受到离群值的左右。（2）零值可能会转换成一个非零值，因此稀疏矩阵将变成一个稠密矩阵。

**MaxAbsScaler**

同样是对某一个特征操作，各特征值除以最大绝对值，因此缩放到[-1,1]之间。且不移动中心点。不会将稀疏矩阵变得稠密。例如一个叫长度的特征，有三个样本有此特征，特征向量为[-1000,100,10],最大绝对值为1000，因此转换为[-1000/1000,100/100,10/1000]=[-1,0.1,0.01]。

因此如果最大绝对值是一个离群点，显然这种处理方式是很不合理的。

**String与Index相互转换**

**VectorIndexer**

主要作用：提高决策树或随机森林等ML方法的分类效果。

VectorIndexer是对数据集特征向量中的类别（离散值）特征（index categorical features categorical features ）进行编号。

它能够自动判断那些特征是离散值型的特征，并对他们进行编号，具体做法是通过设置一个maxCategories，特征向量中某一个特征不重复取值个数小于maxCategories，则被重新编号为0～K（K<=maxCategories-1）。某一个特征不重复取值个数大于maxCategories，则该特征视为连续值，不会重新编号（不会发生任何改变）。结合例子看吧，实在太绕了。

**StringIndexer**

理解了前面的VectorIndexer之后，StringIndexer对数据集的label进行重新编号就很容易理解了，都是采用类似的转换思路，看下面的例子就可以了。

在其它地方应用StringIndexer时还需要注意两个问题：

（1）StringIndexer本质上是对String类型–>index( number);如果是：数值(numeric)–>index(number),实际上是对把数值先进行了类型转换（ cast numeric to string and then index the string values.），也就是说无论是String，还是数值，都可以重新编号（Index）;

（2）利用获得的模型转化新数据集时，可能遇到异常情况，见下面例子。

**IndexToString**

相应的，有StringIndexer，就应该有IndexToString。在应用StringIndexer对labels进行重新编号后，带着这些编号后的label对数据进行了训练，并接着对其他数据进行了预测，得到预测结果，预测结果的label也是重新编号过的，因此需要转换回来。见下面例子，转换回来的convetedPrediction才和原始的label对应。

离散<->连续特征或Label相互转换

**oneHotEncoder**

独热编码将类别特征（离散的，已经转换为数字编号形式），映射成独热编码。这样在诸如Logistic回归这样需要连续数值值作为特征输入的分类器中也可以使用类别（离散）特征。

**Bucketizer**

分箱（分段处理）：将连续数值转换为离散类别

比如特征是年龄，是一个连续数值，需要将其转换为离散类别(未成年人、青年人、中年人、老年人)，就要用到Bucketizer了。

分类的标准是自己定义的，在Spark中为split参数,定义如下：

double[] splits = {0, 18, 35,50， Double.PositiveInfinity}

将数值年龄分为四类0-18，18-35，35-50，55+四个段。

如果左右边界拿不准，就设置为，Double.NegativeInfinity， Double.PositiveInfinity，不会有错的。

**QuantileDiscretizer**

分位树为数离散化，和Bucketizer（分箱处理）一样也是：将连续数值特征转换为离散类别特征。实际上Class QuantileDiscretizer extends （继承自） Class（Bucketizer）。

参数1：不同的是这里不再自己定义splits（分类标准），而是定义分几箱(段)就可以了。QuantileDiscretizer自己调用函数计算分位数，并完成离散化。

* 参数2： 另外一个参数是精度，如果设置为0，则计算最精确的分位数，这是一个高时间代价的操作。

另外上下边界将设置为正负无穷，覆盖所有实数范围。

#### 特征选择

特征选择（Feature Selection）是特征工程中的重要部分，其主要目的是从所有特征中选出相关特征 (relevant feature)和有效特征，也就是说在不引起重要信息丢失的前提下去除掉无关特征 (irrelevant feature) 和冗余特征 (redundant feature)。然后再组成新的、更简单有效的特征两的过程。特征选择是数据分析过程中的常用手段，尤其在高维数据中，可以提出冗余特征，防止过拟合，提升模型性能。进行特征选择的好处主要有以下几种：

1. 降低过拟合风险，提升模型效果

2. 提高训练速度，降低运算开销

3. 更少的特征通常意味着更好的可解释性

过滤式方法中常用的几个方法：卡方检验、F检验和互信息，并探讨它们用于特征选择的内在机理。

既然特征选择的目的是去除无关特征，那么什么是无关特征？ 对于分类问题，在过滤式方法中一般假设与标签独立的特征为无关特征，而卡方检验恰好可以进行独立性检验，所以其适用于特征选择。如果检验结果是某个特征与标签独立，则可以去除该特征。说到卡方检验自然会用到卡方分布，其定义如下[[7]](#footnote-7)：

未完待续。。。

Spark MLlib中实现了三种特征选择的方法，分别是VectorSlicer、RFormula以及ChiSqSelector。

**VectorSlicer**

VectorSlicer是一种转换器，它采用特征向量并输出带有原始特征子数组的新特征向量。这对于从向量列中提取特征很有用。VectorSlicer接受具有指定索引的向量列，然后输出一个新的向量列，其值通过这些索引选择。

索引有两种类型，一是整数索引，代表向量的索引setIndices()。二是字符串索引，表示向量中的要素名称setNames()。这要求向量列具有AttributeGroup，因为实现在Attribute的名称字段上匹配。

以整数和字符串指定都可以接受。此外，您可以同时使用整数索引和字符串名称。必须至少选择一项功能。不允许重复的功能，因此所选索引和名称之间不能有重叠。请注意，如果选择了要素名称，则在遇到空的输入属性时将引发异常。

输出向量将首先按所选索引（按给定顺序）对要素进行排序，然后按所选名称（按给定顺序）对要素进行排序。

假设我们有一个DataFrame与列userFeatures：

userFeatures

[0.0, 10.0, 0.5]

userFeatures是一个包含三个用户特征的向量列。假设userFeature的第一列全部为零，因此我们要删除它并仅选择最后两列。 VectorSlicer使用setIndices（1,2）选择最后两个元素，然后生成一个名为features的新向量列：

|  |  |
| --- | --- |
| userFeatures | features |
| [0.0, 10.0, 0.5] | [10.0, 0.5] |

假设userFeatures有输入属性，如[“f1”，“f2”，“f3”]，那么我们可以使用setNames（“f2”，“f3”）来选择它们。

|  |  |
| --- | --- |
| userFeatures | features |
| [0.0, 10.0, 0.5] | [10.0, 0.5] |
| [”f1”, “f2”, “f3”] | [”f2”, “f3”] |

以下是实现向量选择的一个scala代码示例

import org.apache.spark.sql.Row

import org.apache.spark.sql.types.StructType

val data = Arrays.asList(

Row(Vectors.sparse(3, Seq((0, -2.0), (1, 2.3)))),

Row(Vectors.dense(-2.0, 2.3, 0.0))

)

val defaultAttr = NumericAttribute.defaultAttr

val attrs = Array(“f1”, “f2”, “f3”).map(defaultAttr.withName)

val attrGroup = new AttributeGroup(“userFeatures”, attrs.asInstanceOf[Array[Attribute]])

val dataset = spark.createDataFrame(data, StructType(Array(attrGroup.toStructField())))

val slicer = new VectorSlicer().setInputCol(“userFeatures”).setOutputCol(“features”)

slicer.setIndices(Array(1)).setNames(Array(“f3”))

// or slicer.setIndices(Array(1, 2)), or slicer.setNames(Array(“f2”, “f3”))

val output = slicer.transform(dataset)

output.show(false)

运行结果：

|userFeatures |features |

+--------------------+-------------+

|(3,[0,1],[-2.0,2.3])|(2,[0],[2.3])|

|[-2.0,2.3,0.0] |[2.3,0.0] |

|  |  |
| --- | --- |
|  |  |

**RFormula**

RFormula通过R模型公式来选择列。支持R操作中的部分操作，包括‘~’, ‘.’, ‘:’, ‘+’以及‘-‘，基本操作如下[[8]](#footnote-8)：

1、 ~分隔目标和对象

2、 +合并对象，“+0”意味着删除空格

3、-删除一个对象，“-1”表示删除空格

4、 :交互（数值相乘，类别二值化）

5、 . 除了目标列的全部列

假设a和b为两列：

1、 y ~ a + b表示模型y ~ w0 + w1 \* a +w2 \* b其中w0为截距，w1和w2为相关系数

2、 y ~a + b + a:b – 1表示模型y ~ w1\* a + w2 \* b + w3 \* a \* b，其中w1，w2，w3是相关系数

RFormula产生一个向量特征列以及一个double或者字符串标签列。如果用R进行线性回归，则对String类型的输入列进行one-hot编码、对数值型的输入列进行double类型转化。如果类别列是字符串类型，它将通过StringIndexer转换为double类型。如果标签列不存在，则输出中将通过规定的响应变量创造一个标签列。

示例：假设我们有一个DataFrame含有id,country, hour和clicked四列：

id | country | hour | clicked

---|---------|------|---------

7 | “US” | 18 | 1.0

8 | “CA” | 12 | 0.0

9 | “NZ” | 15 | 0.0

如果我们使用RFormula公式clicked ~ country+ hour，则表明我们希望基于country和hour预测clicked，通过转换我们可以得到如DataFrme：

id | country | hour | clicked | features | label

---|---------|------|---------|------------------|-------

7 | “US” | 18 | 1.0 | [0.0, 0.0, 18.0] | 1.0

8 | “CA” | 12 | 0.0 | [0.0, 1.0, 12.0] | 0.0

9 | “NZ” | 15 | 0.0 | [1.0, 0.0, 15.0] | 0.0

import org.apache.spark.ml.feature.RFormula

val dataset = spark.createDataFrame(Seq(

(7, “US”, 18, 1.0),

(8, “CA”, 12, 0.0),

(9, “NZ”, 15, 0.0)

)).toDF(“id”, “country”, “hour”, “clicked”)

val formula = new RFormula()

.setFormula(“clicked ~ country + hour”)

.setFeaturesCol(“features”)

.setLabelCol(“label”)

val output = formula.fit(dataset).transform(dataset)

output.select(“features”, “label”).show()

**ChiSqSelector**

ChiSqSelector代表卡方特征选择。它适用于带有类别特征的标签数据。ChiSqSelector根据独立卡方检验，然后选取类别标签主要依赖的特征。它类似于选取最有预测能力的特征。它支持三种特征选取方法：

1、numTopFeatures：通过卡方检验选取最具有预测能力的Top(num)个特征；

2、percentile：类似于上一种方法，但是选取一小部分特征而不是固定(num)个特征；

3、fpr:选择P值低于门限值的特征，这样就可以控制false positive rate来进行特征选择；

默认情况下特征选择方法是numTopFeatures(50)，可以根据setSelectorType()选择特征选取方法。

示例：假设我们有一个DataFrame含有id,features和clicked三列，其中clicked为需要预测的目标：

id | features | clicked

---|-----------------------|---------

7 | [0.0, 0.0, 18.0, 1.0] | 1.0

8 | [0.0, 1.0, 12.0, 0.0] | 0.0

9 | [1.0, 0.0, 15.0, 0.1] | 0.0

如果我们使用ChiSqSelector并设置numTopFeatures为1，根据标签clicked，features中最后一列将会是最有用特征：

id | features | clicked | selectedFeatures

---|-----------------------|---------|------------------

7 | [0.0, 0.0, 18.0, 1.0] | 1.0 | [1.0]

8 | [0.0, 1.0, 12.0, 0.0] | 0.0 | [0.0]

9 | [1.0, 0.0, 15.0, 0.1] | 0.0 | [0.1]

import org.apache.spark.ml.feature.ChiSqSelector

import org.apache.spark.ml.linalg.Vectors

val data = Seq(

(7, Vectors.dense(0.0, 0.0, 18.0, 1.0), 1.0),

(8, Vectors.dense(0.0, 1.0, 12.0, 0.0), 0.0),

(9, Vectors.dense(1.0, 0.0, 15.0, 0.1), 0.0)

)

val df = spark.createDataset(data).toDF(“id”, “features”, “clicked”)

val selector = new ChiSqSelector()

.setNumTopFeatures(1)

.setFeaturesCol(“features”)

.setLabelCol(“clicked”)

.setOutputCol(“selectedFeatures”)

val result = selector.fit(df).transform(df)

println(s”ChiSqSelector output with top ${selector.getNumTopFeatures} features selected”)

result.show()

## MongoDB

#### MongoDB介绍

所有案例都采用NoSQL数据库进行存储，不同于传统的关系型数据库管理系统（RDBMS），NoSQL数据库在存储大数据量更加具有优势并且读写性能相对较高，

MongoDB是一个基于分布式文件存储的数据库，由C++语言编写，是一个介于关系数据库和非关系数据库之间的产品。其存储的内容是文档型，数据结构由键值对（key=>value）组成。MOngoDB文档类似于JSON对象，字段值可以包含其他文档，数组和文档数组。

比如：

图片包含 物体

描述已自动生成

#### mongodb 4.0安装部署

下载安装包，并解压tgz文件：

$ curl -O <https://fastdl.mongodb.org/linux/mongodb-linux-x86_64-4.0.5.tgz>

$ tar -xvf mongodb-linux-x86\_64-4.0.5.tgz

将压缩包拷贝到指定文件夹（/opt）

$ mv mongodb-linux-x86\_64-4.0.5 /opt/mongodb

添加mongodb的环境变量，并bin目录添加到PATH路径中：

vim /home/hadoop/.bashrc

export MONGODB\_HOME=/opt/mongodb

export PATH=$PATH:$MONGODB\_HOME/bin

source /home/hadoop/.bashrc

打开 /opt/mongodb/bin/mongodb.conf配置文件添加配置项：

#数据文件目录（寻找一个稍大的磁盘，如/mnt/disk0/）

dbpath = /mnt/disk0/mongodb/data/db

#日志目录

logpath = /mnt/disk0/mongodb/logs/mongodb.log

#端口

port = 27017

#以守护进程运行

fork = true

#设置权限

auth=True

bind\_ip=127.0.0.1

根据上述配置文件创建数据库目录：

$ mkdir -p /mnt/disk0/mongodb/data/db

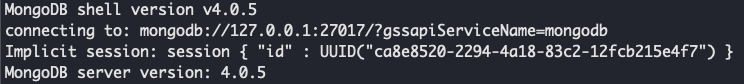
$ mkdir -p /mnt/disk0/mongodb/logs/

**运行mongodb：**

$ /opt/mongodb/bin/mongod -f /opt/mongodb/bin/mongodb.conf

测试mongodb服务：

$ mongo



**创建管理员：**

* db.createUser({user: ‘root’, pwd: ‘dis2019’, roles: [’root’]})

给recommender设置权限：

* db.createUser({user:’admin’,pwd:’dis2019’,roles: [{role:’readWrite’,db:’recommender’}]}))验证：
* db.auth(‘admin’, ‘dis2019’)

普通db用户：

* db.createUser({user: “admin”, pwd: “dis2019”, roles: [{ role: “dbOwner”, db: “recommender” }]})
* db.auth(‘admin’, ‘dis2019’)

重启mongo，登录 ：

mongo -u admin -p ‘dis2019’—authenticationDatabase “recommender”

或者，命令行登录方式

mongo 127.0.0.1:27017/recommender -u admin -p dis2019

#### mongodb的基础命令

查看collection：

show collections

创建collection：

db.createCollection(“u\_attr”)

#### spark读写MongoDB

为了提升推荐系统的实时分析功能，集成Spark引擎可以扩展MongoDB对于数据的处理和运营。 由此，可以使用集成工具MongoDB Connector for Apache Spark，这让Spark操作MongoDB数据更为简单，方便Spark进行流失数据处理。

添加sbt依赖包，在build.sbt中添加：

libraryDependencies += “org.mongodb.spark” %% “mongo-spark-connector” % sparkVersion

要求MongoDB版本在2.6及以上， Spark版本在2.4.0及以上，Scala版本在2.11及以上。

导入MongoDB连接包：

import com.mongodb.spark.\_

连接到MongoDB

当RDD的action操作需要读写MongoDB时，MongoDB数据库连接会自动触发。

从MongoDB中读取数据

import com.mongodb.spark.config.\_

def sparkReadCol(spark: SparkSession,  
 mongoUri: String,  
 colName:String)={  
 val readConfig = ReadConfig(Map(  
 “uri” -> mongoUri,  
 “collection” -> colName  
 ))  
 MongoSpark.load(spark.sparkContext, readConfig)  
}

#调用MongoDB的读取方法

val mongoDBName = “recommender”  
var mongoColName = “spark”  
val mongoUser = “admin”  
val mongoPwd = “admin”  
val mongoUriStr = s”mongodb://${mongoUser}:${mongoPwd}@127.0.0.1:27017/${mongoDBName}”

val doc = sparkReadCol(spark, mongoUriStr, mongoColName)

println(doc.count)

println(doc.first.toJson)

RDD数据写入MongoDB：

import com.mongodb.spark.config.\_

def sparkWriteCol(sparkDocuments: RDD[Document],  
 mongoUri: String,  
 colName:String) ={  
 val writeConfig = WriteConfig(Map(  
 “uri” -> mongoUri,  
 “collection” -> colName  
 ))  
 MongoSpark.save(u\_result, writeConfig)  
}

val writeConfig = WriteConfig(Map(“collection” -> “ recommender”, “writeConcern” -> “testing”), Some(WriteConfig(sc)))

val sparkDocuments = sc.parallelize((1 to 10).map(i => Document.parse(s”{spark: $i}”)))

val mongoDBName = “recommender”  
var mongoColName = “spark”  
val mongoUser = “admin”  
val mongoPwd = “admin”  
val mongoUriStr = s”mongodb://${mongoUser}:${mongoPwd}@127.0.0.1:27017/${mongoDBName}”

#调用写入方法

sparkWriteCol(sparkDocuments, mongoUriStr, mongoColName)

## **Kafka**

#### Kafka消息队列

Kafka最初是由LinkedIn开发的一个分布式的消息系统，使用Scala编写，其代码于2011年提交Apache软件基金会开源。Kafka为实时数据提供了一个统一、高吞吐和低延迟的平台，由此被广泛使用。目前越来越多的开源分布式处理系统如Spark、Storm、Flink都支持与Kafka集成。

Kafka存储的消息来自任意多被称为“生产者”（Producer）的进程。数据从而可以被分配到不同的“分区”（Partition）、不同的“Topic”下。在一个分区内，这些消息被索引并连同时间戳存储在一起。其它被称为“消费者”（Consumer）的进程可以从分区查询消息。Kafka运行在一个由一台或多台服务器组成的集群上，并且分区可以跨集群结点分布[[9]](#footnote-9)。

Kafka高效地处理实时流式数据，可以实现与Storm、HBase和Spark的集成。作为聚类部署到多台服务器上，Kafka处理它所有的发布和订阅消息系统使用了四个API，即生产者API、消费者API、Stream API和Connector API。它能够传递大规模流式消息，自带容错功能，已经取代了一些传统消息系统，如JMS、AMQP等。

Kafka架构的主要术语包括Topic、Record和Broker。Topic由Record组成，Record持有不同的信息，而Broker则负责复制消息。Kafka有四个主要API：

生产者API：支持应用程序发布Record流。

消费者API：支持应用程序订阅Topic和处理Record流。

Stream API：将输入流转换为输出流，并产生结果。

Connector API：执行可重用的生产者和消费者API，可将Topic链接到现有应用程序。

#### **kafka安装（同时启动zookeeper）**：

最新版本下载

<http://mirrors.tuna.tsinghua.edu.cn/apache/kafka/2.1.0/kafka_2.11-2.1.0.tgz>

修改配置文件

server.properties：

标识当前server在集群中id，从0开始，每台server必须集群内唯一

broker.id=0

port=9092

host.name=davinci01

log.dirs最好匹配多个磁盘上，以便提高读写效率

log.dirs=/opt/kafka/logs

这些目录必须已经手动创建好了

Zookeeper配置，2181端口为默认端口

zookeeper.connect=localhost:2181

日志中每个segment的大小为1G

log.segment.bytes=1073741824

其他的不变

scp到各节点

在/opt/kafka目录下建立logs目录以存放server日志

启动：

nohup ./bin/zookeeper-server-start.sh config/zookeeper.properties > logs/zookeeper.log &

nohup ./bin/kafka-server-start.sh config/server.properties > logs/server.log &

创建topic

bin/kafka-topics.sh -zookeeper jh-a0-010:2181 -topic u\_sim -replication-factor 1 -partitions 2 -create

查看topic状态

/opt/kafka/bin/kafka-topics.sh -zookeeper jh-a0-010:2181/kafka -list

创建producer

bin/kafka-console-producer.sh—broker-list jh-a0-010:9092 --topic u\_sim

在producer测输入：

hello kafka

在consumer观察：

bin/kafka-console-consumer.sh—bootstrap-server jh-a0-010:9092 --topic u\_sim—from-beginning

**kafka单机测试：**

注意： 单机配置中host为localhost, zookeeper设置为2181/kakfa，以后需要重新改

/opt/kafka/bin/kafka-console-producer.sh—broker-list localhost:9092 --topic u\_sim

/opt/kafka/bin/kafka-console-consumer.sh—bootstrap-server localhost:9092 --topic u\_sim

## Redis

#### Redis简介

Redis是一个key-value格式的内存存储数据库，读写速度很快，并且支持丰富的数据结果，包括以下：

1.String

2. list

3. hash table

4. set(集合)

5. zset(有序集合)

redis的体量很小，但却支持丰富的数据结构和相应的操作 方式，是一个非常好的工具。

#### 安装

wget <http://download.redis.io/releases/redis-5.0.3.tar.gz>

$ tar xzf redis-5.0.3.tar.gz

$cd redis-5.0.3

$ make

运行：

src/redis-server

客户端：

/opt/redis/src/redis-cli—raw

#如果需要密码

* auth “dis2019”
* set foo bar

OK

* get foo
* keys \* //查询所有key
* dbsize //查询key数目

配置： 配置文件启动未完成

#### Redis数据结构[[10]](#footnote-10)

下面介绍这几种常见的数据结构和使用方法：

1.String

常用命令: set,get,decr,incr,mget等。

String数据结构是简单的key-value类型，value其实不仅可以是String，也可以是数字。 常规key-value缓存应用； 常规计数：微博数，粉丝数等。

2.Hash

常用命令： hget,hset,hgetall等。

hash是一个string类型的field和value的映射表，hash特别适合用于存储对象，后续操作的时候，你可以直接仅仅修改这个对象中的某个字段的值。 比如我们可以hash数据结构来存储用户信息，商品信息等等。比如下面我就用hash类型存放了我本人的一些信息：

key=JavaUser293847 value={ “id”: 1, “name”: “SnailClimb”, “age”: 22, “location”: “Wuhan, Hubei” }

3.List

常用命令: lpush,rpush,lpop,rpop,lrange等

list就是链表，Redis list的应用场景非常多，也是Redis最重要的数据结构之一，比如微博的关注列表，粉丝列表，消息列表等功能都可以用Redis的list结构来实现。

Redis list的实现为一个双向链表，即可以支持反向查找和遍历，更方便操作，不过带来了部分额外的内存开销。

另外可以通过lrange命令，就是从某个元素开始读取多少个元素，可以基于list实现分页查询，这个很棒的一个功能，基于redis实现简单的高性能分页，可以做类似微博那种下拉不断分页的东西（一页一页的往下走），性能高。

4.Set

常用命令： sadd,spop,smembers,sunion等

set对外提供的功能与list类似是一个列表的功能，特殊之处在于set是可以自动排重的。

当你需要存储一个列表数据，又不希望出现重复数据时，set是一个很好的选择，并且set提供了判断某个成员是否在一个set集合内的重要接口，这个也是list所不能提供的。可以基于set轻易实现交集、并集、差集的操作。

比如：在微博应用中，可以将一个用户所有的关注人存在一个集合中，将其所有粉丝存在一个集合。Redis可以非常方便的实现如共同关注、共同粉丝、共同喜好等功能。这个过程也就是求交集的过程，具体命令如下：

sinterstore key1 key2 key3将交集存在key1内

5.Sorted Set

常用命令： zadd,zrange,zrem,zcard等

和set相比，sorted set增加了一个权重参数score，使得集合中的元素能够按score进行有序排列。

举例： 在直播系统中，实时排行信息包含直播间在线用户列表，各种礼物排行榜，弹幕消息（可以理解为按消息维度的消息排行榜）等信息，适合使用Redis中的Sorted Set结构进行存储。

**redis设置过期时间**

Redis中有个设置时间过期的功能，即对存储在redis数据库中的值可以设置一个过期时间。作为一个缓存数据库，这是非常实用的。如我们一般项目中的token或者一些登录信息，尤其是短信验证码都是有时间限制的，按照传统的数据库处理方式，一般都是自己判断过期，这样无疑会严重影响项目性能。

我们set key的时候，都可以给一个expire time，就是过期时间，通过过期时间我们可以指定这个key可以存活的时间。

如果假设你设置了一批key只能存活1个小时，那么接下来1小时后，redis是怎么对这批key进行删除的？

定期删除+惰性删除。

通过名字大概就能猜出这两个删除方式的意思了。

定期删除：redis默认是每隔100ms就随机抽取一些设置了过期时间的key，检查其是否过期，如果过期就删除。注意这里是随机抽取的。为什么要随机呢？你想一想假如redis存了几十万个key ，每隔100ms就遍历所有的设置过期时间的key的话，就会给CPU带来很大的负载！

惰性删除 ：定期删除可能会导致很多过期key到了时间并没有被删除掉。所以就有了惰性删除。假如你的过期key，靠定期删除没有被删除掉，还停留在内存里，除非你的系统去查一下那个key，才会被redis给删除掉。这就是所谓的惰性删除，也是够懒的哈！

但是仅仅通过设置过期时间还是有问题的。我们想一下：如果定期删除漏掉了很多过期key，然后你也没及时去查，也就没走惰性删除，此时会怎么样？如果大量过期key堆积在内存里，导致redis内存块耗尽了。怎么解决这个问题呢？ redis内存淘汰机制。

redis内存淘汰机制(MySQL里有2000w数据，Redis中只存20w的数据，如何保证Redis中的数据都是热点数据?)

redis配置文件redis.conf中有相关注释，我这里就不贴了，大家可以自行查阅或者通过这个网址查看：

<http://download.redis.io/redis-stable/redis.conf>

redis提供6种数据淘汰策略：

volatile-lru：从已设置过期时间的数据集（server.db[i].expires）中挑选最近最少使用的数据淘汰

volatile-ttl：从已设置过期时间的数据集（server.db[i].expires）中挑选将要过期的数据淘汰

volatile-random：从已设置过期时间的数据集（server.db[i].expires）中任意选择数据淘汰

allkeys-lru：当内存不足以容纳新写入数据时，在键空间中，移除最近最少使用的key（这个是最常用的）

allkeys-random：从数据集（server.db[i].dict）中任意选择数据淘汰

no-eviction：禁止驱逐数据，也就是说当内存不足以容纳新写入数据时，新写入操作会报错。这个应该没人使用吧！

**4.0版本后增加以下两种：**

volatile-lfu：从已设置过期时间的数据集(server.db[i].expires)中挑选最不经常使用的数据淘汰

allkeys-lfu：当内存不足以容纳新写入数据时，在键空间中，移除最不经常使用的key

备注： 关于redis设置过期时间以及内存淘汰机制，我这里只是简单的总结一下，后面会专门写一篇文章来总结！

redis持久化机制

(怎么保证redis挂掉之后再重启数据可以进行恢复)

很多时候我们需要持久化数据也就是将内存中的数据写入到硬盘里面，大部分原因是为了之后重用数据（比如重启机器、机器故障之后恢复数据），或者是为了防止系统故障而将数据备份到一个远程位置。

Redis不同于Memcached的很重一点就是，Redis支持持久化，而且支持两种不同的持久化操作。Redis的一种持久化方式叫快照（snapshotting，RDB），另一种方式是只追加文件（append-only file,AOF）。这两种方法各有千秋，下面我会详细这两种持久化方法是什么，怎么用，如何选择适合自己的持久化方法。

**redis使用注意点[[11]](#footnote-11)：**

1. 管理好自己的key，比如用一个set来存放所有同类型的key，这样可以避免使用keys来获取该类型所有key，由于Redis是单线程的，在数据量大的情况下将会阻塞Redis中其他的操作

2. 如果不能避免使用keys操作，那么使用scan来替代，将遍历Redis DB中所有key的操作放到客户端来做，这样就不会导致其他操作被阻塞了

3. 在数据量大的情况下，key或value的微小的压缩也能带来不小的内存使用提高

4. hash结构是个好东西，但是它只支持一级hash，所以如果需要使用多级hash，那就进行序列化吧

5. 注意客户端和服务端之间的网络通信情况(–latency/–latency-history)

6. Redis的python接口返回的key的类型和value中的最小结构的类型都是string，记得转换成自己需要的类型

## ElasticSearch

#### ES简介

#### ES安装[[12]](#footnote-12)

Elasticsearch依赖Java8环境，先安装JDK并配置JAVA\_HOME环境变量。

安装完 Java，就可以跟着官方文档安装 Elastic。直接下载压缩包比较简单。

$ wget https://artifacts.elastic.co/downloads/elasticsearch/elasticsearch-5.5.1.zip

$ unzip elasticsearch-5.5.1.zip

$ cd elasticsearch-5.5.1/

接着，进入解压后的目录，运行下面的命令，启动 Elastic。

$ ./bin/elasticsearch

如果这时报错"max virtual memory areas vm.maxmapcount [65530] is too low"，要运行下面的命令。

$ sudo sysctl -w vm.max\_map\_count=262144

如果一切正常，Elastic 就会在默认的9200端口运行。这时，打开另一个命令行窗口，请求该端口，会得到说明信息。

$ curl localhost:9200

{

"name" : "atntrTf",

"cluster\_name" : "elasticsearch",

"cluster\_uuid" : "tf9250XhQ6ee4h7YI11anA",

"version" : {

"number" : "5.5.1",

"build\_hash" : "19c13d0",

"build\_date" : "2017-07-18T20:44:24.823Z",

"build\_snapshot" : false,

"lucene\_version" : "6.6.0"

},

"tagline" : "You Know, for Search"

}

上面代码中，请求9200端口，Elastic 返回一个 JSON 对象，包含当前节点、集群、版本等信息。

按下 Ctrl + C，Elastic 就会停止运行。

默认情况下，Elastic 只允许本机访问，如果需要远程访问，可以修改 Elastic 安装目录的config/elasticsearch.yml文件，去掉network.host的注释，将它的值改成0.0.0.0，然后重新启动 Elastic。

network.host: 0.0.0.0

上面代码中，设成0.0.0.0让任何人都可以访问。线上服务不要这样设置，要设成具体的 IP。

#### 基本概念

2.1 Node 与 Cluster

Elastic 本质上是一个分布式数据库，允许多台服务器协同工作，每台服务器可以运行多个 Elastic 实例。

单个 Elastic 实例称为一个节点（node）。一组节点构成一个集群（cluster）。

2.2 Index

Elastic 会索引所有字段，经过处理后写入一个反向索引（Inverted Index）。查找数据的时候，直接查找该索引。

所以，Elastic 数据管理的顶层单位就叫做 Index（索引）。它是单个数据库的同义词。每个 Index （即数据库）的名字必须是小写。

下面的命令可以查看当前节点的所有 Index。

$ curl -X GET 'http://localhost:9200/\_cat/indices?v'

2.3 Document

Index 里面单条的记录称为 Document（文档）。许多条 Document 构成了一个 Index。

Document 使用 JSON 格式表示，下面是一个例子。

{

"user": "张三",

"title": "工程师",

"desc": "数据库管理"

}

同一个 Index 里面的 Document，不要求有相同的结构（scheme），但是最好保持相同，这样有利于提高搜索效率。

2.4 Type

Document 可以分组，比如weather这个 Index 里面，可以按城市分组（北京和上海），也可以按气候分组（晴天和雨天）。这种分组就叫做 Type，它是虚拟的逻辑分组，用来过滤 Document。

不同的 Type 应该有相似的结构（schema），举例来说，id字段不能在这个组是字符串，在另一个组是数值。这是与关系型数据库的表的一个区别。性质完全不同的数据（比如products和logs）应该存成两个 Index，而不是一个 Index 里面的两个 Type（虽然可以做到）。

下面的命令可以列出每个 Index 所包含的 Type。

$ curl 'localhost:9200/\_mapping?pretty=true'

根据规划，Elastic 6.x 版只允许每个 Index 包含一个 Type，7.x 版将会彻底移除 Type。

#### 新建和删除 Index

新建 Index，可以直接向 Elastic 服务器发出 PUT 请求。下面的例子是新建一个名叫weather的 Index。

$ curl -X PUT 'localhost:9200/weather'

服务器返回一个 JSON 对象，里面的acknowledged字段表示操作成功。

{

"acknowledged":true,

"shards\_acknowledged":true

}

然后，我们发出 DELETE 请求，删除这个 Index。

$ curl -X DELETE 'localhost:9200/weather'

四、中文分词设置

首先，安装中文分词插件。这里使用的是 ik，也可以考虑其他插件（比如 smartcn）。

$ ./bin/elasticsearch-plugin install https://github.com/medcl/elasticsearch-analysis-ik/releases/download/v5.5.1/elasticsearch-analysis-ik-5.5.1.zip

上面代码安装的是5.5.1版的插件，与 Elastic 5.5.1 配合使用。

接着，重新启动 Elastic，就会自动加载这个新安装的插件。

然后，新建一个 Index，指定需要分词的字段。这一步根据数据结构而异，下面的命令只针对本文。基本上，凡是需要搜索的中文字段，都要单独设置一下。

$ curl -X PUT 'localhost:9200/accounts' -d '

{

"mappings": {

"person": {

"properties": {

"user": {

"type": "text",

"analyzer": "ik\_max\_word",

"search\_analyzer": "ik\_max\_word"

},

"title": {

"type": "text",

"analyzer": "ik\_max\_word",

"search\_analyzer": "ik\_max\_word"

},

"desc": {

"type": "text",

"analyzer": "ik\_max\_word",

"search\_analyzer": "ik\_max\_word"

}

}

}

}

}'

上面代码中，首先新建一个名称为accounts的 Index，里面有一个名称为person的 Type。person有三个字段。

user

title

desc

这三个字段都是中文，而且类型都是文本（text），所以需要指定中文分词器，不能使用默认的英文分词器。

Elastic 的分词器称为 analyzer。我们对每个字段指定分词器。

"user": {

"type": "text",

"analyzer": "ik\_max\_word",

"search\_analyzer": "ik\_max\_word"

}

上面代码中，analyzer是字段文本的分词器，search\_analyzer是搜索词的分词器。ik\_max\_word分词器是插件ik提供的，可以对文本进行最大数量的分词。

#### 数据操作

5.1 新增记录

向指定的 /Index/Type 发送 PUT 请求，就可以在 Index 里面新增一条记录。比如，向/accounts/person发送请求，就可以新增一条人员记录。

$ curl -X PUT 'localhost:9200/accounts/person/1' -d '

{

"user": "张三",

"title": "工程师",

"desc": "数据库管理"

}'

服务器返回的 JSON 对象，会给出 Index、Type、Id、Version 等信息。

{

"\_index":"accounts",

"\_type":"person",

"\_id":"1",

"\_version":1,

"result":"created",

"\_shards":{"total":2,"successful":1,"failed":0},

"created":true

}

如果你仔细看，会发现请求路径是/accounts/person/1，最后的1是该条记录的 Id。它不一定是数字，任意字符串（比如abc）都可以。

新增记录的时候，也可以不指定 Id，这时要改成 POST 请求。

$ curl -X POST 'localhost:9200/accounts/person' -d '

{

"user": "李四",

"title": "工程师",

"desc": "系统管理"

}'

上面代码中，向/accounts/person发出一个 POST 请求，添加一个记录。这时，服务器返回的 JSON 对象里面，\_id字段就是一个随机字符串。

{

"\_index":"accounts",

"\_type":"person",

"\_id":"AV3qGfrC6jMbsbXb6k1p",

"\_version":1,

"result":"created",

"\_shards":{"total":2,"successful":1,"failed":0},

"created":true

}

注意，如果没有先创建 Index（这个例子是accounts），直接执行上面的命令，Elastic 也不会报错，而是直接生成指定的 Index。所以，打字的时候要小心，不要写错 Index 的名称。

5.2 查看记录

向/Index/Type/Id发出 GET 请求，就可以查看这条记录。

$ curl 'localhost:9200/accounts/person/1?pretty=true'

上面代码请求查看/accounts/person/1这条记录，URL 的参数pretty=true表示以易读的格式返回。

返回的数据中，found字段表示查询成功，\_source字段返回原始记录。

{

"\_index" : "accounts",

"\_type" : "person",

"\_id" : "1",

"\_version" : 1,

"found" : true,

"\_source" : {

"user" : "张三",

"title" : "工程师",

"desc" : "数据库管理"

}

}

如果 Id 不正确，就查不到数据，found字段就是false。

$ curl 'localhost:9200/weather/beijing/abc?pretty=true'

{

"\_index" : "accounts",

"\_type" : "person",

"\_id" : "abc",

"found" : false

}

5.3 删除记录

删除记录就是发出 DELETE 请求。

$ curl -X DELETE 'localhost:9200/accounts/person/1'

这里先不要删除这条记录，后面还要用到。

5.4 更新记录

更新记录就是使用 PUT 请求，重新发送一次数据。

$ curl -X PUT 'localhost:9200/accounts/person/1' -d '

{

"user" : "张三",

"title" : "工程师",

"desc" : "数据库管理，软件开发"

}'

{

"\_index":"accounts",

"\_type":"person",

"\_id":"1",

"\_version":2,

"result":"updated",

"\_shards":{"total":2,"successful":1,"failed":0},

"created":false

}

上面代码中，我们将原始数据从"数据库管理"改成"数据库管理，软件开发"。 返回结果里面，有几个字段发生了变化。

"\_version" : 2,

"result" : "updated",

"created" : false

可以看到，记录的 Id 没变，但是版本（version）从1变成2，操作类型（result）从created变成updated，created字段变成false，因为这次不是新建记录。

#### 数据查询

6.1 返回所有记录

使用 GET 方法，直接请求/Index/Type/\_search，就会返回所有记录。

$ curl 'localhost:9200/accounts/person/\_search'

{

"took":2,

"timed\_out":false,

"\_shards":{"total":5,"successful":5,"failed":0},

"hits":{

"total":2,

"max\_score":1.0,

"hits":[

{

"\_index":"accounts",

"\_type":"person",

"\_id":"AV3qGfrC6jMbsbXb6k1p",

"\_score":1.0,

"\_source": {

"user": "李四",

"title": "工程师",

"desc": "系统管理"

}

},

{

"\_index":"accounts",

"\_type":"person",

"\_id":"1",

"\_score":1.0,

"\_source": {

"user" : "张三",

"title" : "工程师",

"desc" : "数据库管理，软件开发"

}

}

]

}

}

上面代码中，返回结果的 took字段表示该操作的耗时（单位为毫秒），timed\_out字段表示是否超时，hits字段表示命中的记录，里面子字段的含义如下。

total：返回记录数，本例是2条。

max\_score：最高的匹配程度，本例是1.0。

hits：返回的记录组成的数组。

返回的记录中，每条记录都有一个\_score字段，表示匹配的程序，默认是按照这个字段降序排列。

6.2 全文搜索

Elastic 的查询非常特别，使用自己的查询语法，要求 GET 请求带有数据体。

$ curl 'localhost:9200/accounts/person/\_search' -d '

{

"query" : { "match" : { "desc" : "软件" }}

}'

上面代码使用 Match 查询，指定的匹配条件是desc字段里面包含"软件"这个词。返回结果如下。

{

"took":3,

"timed\_out":false,

"\_shards":{"total":5,"successful":5,"failed":0},

"hits":{

"total":1,

"max\_score":0.28582606,

"hits":[

{

"\_index":"accounts",

"\_type":"person",

"\_id":"1",

"\_score":0.28582606,

"\_source": {

"user" : "张三",

"title" : "工程师",

"desc" : "数据库管理，软件开发"

}

}

]

}

}

Elastic 默认一次返回10条结果，可以通过size字段改变这个设置。

$ curl 'localhost:9200/accounts/person/\_search' -d '

{

"query" : { "match" : { "desc" : "管理" }},

"size": 1

}'

上面代码指定，每次只返回一条结果。

还可以通过from字段，指定位移。

$ curl 'localhost:9200/accounts/person/\_search' -d '

{

"query" : { "match" : { "desc" : "管理" }},

"from": 1,

"size": 1

}'

上面代码指定，从位置1开始（默认是从位置0开始），只返回一条结果。

6.3 逻辑运算

如果有多个搜索关键字， Elastic 认为它们是or关系。

$ curl 'localhost:9200/accounts/person/\_search' -d '

{

"query" : { "match" : { "desc" : "软件 系统" }}

}'

上面代码搜索的是软件 or 系统。

如果要执行多个关键词的and搜索，必须使用布尔查询。

$ curl 'localhost:9200/accounts/person/\_search' -d '

{

"query": {

"bool": {

"must": [

{ "match": { "desc": "软件" } },

{ "match": { "desc": "系统" } }

]

}

}

}'

第二部分 推荐系统相关算法

# 大规模相似度计算

## 为什么计算相似度

如果你想找出不同用户和不同内容之间的相似之处，就必须进行相似度计算。通过相似度计算，我们可以找到跟我喜欢的物品相似的其他物品，对应的，也可以找到跟我志趣相投的其他用户（因为我喜欢的物品，他们也喜欢）。

那么如何定义相似度？如果我们将两个用户的相似度范围限定在0-1，0是完全不相似度，1为最相似。接下来，需要限定相似度的计算场景，比如在电影推荐中，用户之间的相似度指的是用户对于电影的口味或者是兴趣。比如两个用户都喜欢科幻电影，或者都喜欢某个明星参演的电影。但是，我们知道用户的兴趣是多元化的，喜欢科幻电影的人也可能会喜欢其他口味的电影类型。所以，加入其他补充信息，比如用户的年龄、性别等基础人口统计信息，得到的相似度结果会更加合理。

在数学上通常将相似度计算归结为距离计算，也就是将两个用户或者项目抽象为空间中的两个点，然后计算两个点之间的距离。距离越短，相似度越高。

## 相似度的计算方式

距离度量和相似度度量[[13]](#footnote-13)：

距离度量是计算两个对象之间的差异程度。

相似度量是计算两个对象之间的近似程度。

未完待续。。。

相似度的计算方式有很多，每一种适合的场景也不一样。下面介绍几种常用的相似度计算方法：

#### 杰卡尔德距离（Jaccard Distance）

1. 杰卡德相似系数  
   两个集合A和B的交集元素在A，B的并集中所占的比例，称为两个集合的杰卡德相似系数，用符号J(A,B)表示。

(2) 杰卡德距离  
 与杰卡德相似系数相反的概念是杰卡德距离(Jaccard distance)。杰卡德距离可用如下公式表示：

杰卡德距离用两个集合中不同元素占所有元素的比例来衡量两个集合的区分度。

假设有一个矩阵，横向代表item，包括row1、row2和row3，纵向代表user的购买情况，如果买过某个item，则对应数值为1，反之则为0。

user

item

通过python计算item两两之间的杰卡尔德相似度：

import scipy.spatial.distance as dist

v1=np.array([0,1,0,1,0,1])

v2=np.array([0,1,1,0,1,0])

v3=np.array([0,1,1,0,5,0])

matrix1=np.array([v1,v2])

matrix2=np.array([v1,v3])

print(matrix1)

print(matrix2)

ds1=dist.pdist(matrix1, ‘jaccard’)

ds2=dist.pdist(matrix2, ‘jaccard’)

print(ds1)

print(ds2)

其输出为：

[0.8]

[0,8]

上面代码中，v3相较于v2改变了第5个元素的评分，但是v1、v2和v1、v3之间的相似度保持不变。所以，杰卡尔德相似度只会判断是否购买，并不会计算具体评分。

#### 欧式距离（Euclidean Distance）

欧式距离是比较容易理解的一种距离计算方法，来源于空间中两点的距离公式。用向量表示两个n维向量与之间的欧式距离为：

假设一个矩阵，横向代表item，包括row1、row2和row3，纵向代表user的购买情况。矩阵中各个元素是user对item的具体评分，实际场景中评分是Double类型。

user

item

使用python计算row1和row2之间的欧式距离：

vector1 = np.array([0,2,0,4,0,1])

vector2 = np.array([0,1,1,4,1,0])

vector3 = np.array([0,1,3,4,1,0])

vector4 = np.array([1,0,1,4,1,0])

def EuclideanDistance(vector1, vector2):  
 #ds1=np.sqrt(np.sum(np.square(vector1-vector2)))   
 # np.linalg.norm用于范数计算，默认是二范数，相当于平方和开根号  
 return 1.0/(1.0 + np.linalg.norm(vector1 - vector2))

ds1= EuclideanDistance(vector1, vector2)

ds2= EuclideanDistance(vector1, vector3)

ds3= EuclideanDistance(vector1, vector4)

**print**(ds1)

**print**(ds2)

print(ds3)

其输出为：

0.3333333333333333

0.2240092377397959

0.2612038749637414

#### 余弦相似度（Cosine）

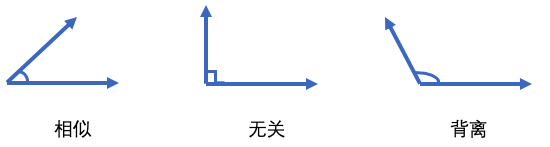
几何中的夹角余弦可以用来衡量两个向量方向之间的差异，放在推荐系统中，可以用来衡量user或者item的行为向量之间的差异。两个n维样本点与用类似于夹角余弦的概念来衡量他们之间的相似度。

假设用X和用户X的向量分别如下：

那么用户X和用户Y的余弦相似度为：

其中表示两个向量的内积，是向量的模，sim(X,Y)就是相似度的值。

从下图简单示意中看出，夹角余弦就是用两个向量之间的夹角大小对应相似的程度。夹角越小代表越相似。如果两个向量则正交余弦值为0，如果夹角为钝角则相似度为负值。



用python计算下列矩阵中row1和row2的相似度：

user

item

vector1 = np.array([0,2,0,2,0,2])

vector2 = np.array([0,1,1,4,1,0])

vector3 = np.array([0,1,3,4,1,0])

vector4 = np.array([0,4,0,4,0,4])

def cosine(vector1, vector2):

cos=np.dot(vector1,vector2)/(np.linalg.norm(vector1)\*(np.linalg.norm(vector2)))

return cos

cos1=cosine(vector1,vector2)

cos2=cosine(vector1,vector3)

cos3=cosine(vector1,vector4)

print(cos1)

print(cos2)

print(cos3)

输出为：

0.6622661785325219

0.5555555555555556

1.0000000000000002

#### 皮尔逊相关系数（Pearson Correlation Coefficient）

皮尔逊相关系数公式实际上就是在计算夹角余弦之前将两个向量减去各个样本的平均值，达到中心化的目的。从知友的回答可以明白，皮尔逊相关函数是余弦相似度在维度缺失上面的一种改进方法。

利用pytho-numpy实现皮尔逊相关系数计算：

def Pearson(dataA,dataB):

# 皮尔逊相关系数的取值范围(-1 ~ 1),0.5 + 0.5 \* result归一化(0 ~ 1)

return 0.5 + 0.5 \* np.corrcoef(dataA,dataB,rowvar = 0)[0][1]

#### 修正余弦相似度（Adjusted Cosine）

为什么需要在余弦相似度的基础上使用修正余弦相似度。因为余弦相似度更多的是从方向上区分向量差异，而对绝对的数值不敏感，因此没法衡量每个维度上数值的差异，会导致一些问题。比如在推荐系统中，X和Y两个用户对两个内容的评分分别为（1,2）和（4,5），使用余弦相似度得到的结果是0.98，两者极为相似。但从评分上看X似乎不喜欢2这个内容，而Y则比较喜欢，余弦相似度对数值的不敏感导致了结果的误差，需要修正这种不合理性就出现了修正余弦相似度。

修正余弦的做法是在所有维度上的数值都减去一个均值，比如X和Y的评分均值为3，那么调整之后为（-2，-1）和（1，2），再使用余弦相似度计算得到-0.8，相似度为复制并且差异不小，不过显然更加符合现实情况。

利用python实现修正余弦相似度：

# 修正cosine减去的是对item i打过分的每个user u，其打分的均值

data = np.mat([[1,2,3],[3,4,5]])

avg = np.mean(data[:,0]) # 下标0表示正在打分的用户

def AdjustedCosine(dataA,dataB,avg):

sumData = (dataA - avg) \* (dataB - avg).T # 若列为向量则为dataA.T \* dataB

denom = np.linalg.norm(dataA - avg) \* np.linalg.norm(dataB - avg)

return 0.5 + 0.5 \* (sumData / denom)

print(AdjustedCosine(data[0,:],data[1,:],avg))

#### 汉明距离（Hamming Distance）

两个等长字符串s1与s2之间的汉明距离定义为将其中一个变为另外一个所需要作的最小替换次数。例如字符串“1111”与“1001”之间的汉明距离为2。

v1=np.array([1,1,0,1,0,1,0,0,1])

v2=np.array([0,1,1,0,0,0,1,1,1])

smstr=np.nonzero(v1-v2)

print(smstr) # 不为0的元素的下标

sm= np.shape(smstr[0])[0]

print( sm )

#输出

#(array([0, 2, 3, 5, 6, 7]),)

#6

#### 曼哈顿距离（Manhatten Distance）

从名字就可以猜出这种距离的计算方法了。想象你在曼哈顿要从一个十字路口开车到另外一个十字路口，驾驶距离是两点间的直线距离吗？显然不是，除非你能穿越大楼。实际驾驶距离就是这个“曼哈顿距离”(L1范数)。而这也是曼哈顿距离名称的来源，曼哈顿距离也称为城市街区距离(City Block distance)。

两个n维向量与之间的曼哈顿距离为：

利用python代码计算曼哈顿距离：

vector1 = np.array([1,2,3])

vector2 = np.array([4,5,6])

op3=np.sum(np.abs(vector1-vector2))

op4=np.linalg.norm(vector1-vector2,ord=1)

#输出

#9

#9.0

#### 对比分析

**杰卡尔德相似度适用场景**

简单来看，每一个相似度度量指标都不是完美的，比如杰卡尔德相似度无法将具体评分带入公式中，只是判断有和没有。所以，一般在不考虑项目的具体数值时使用。

**欧式距离相似度适用场景**

欧式距离在几何意义上，是对向量距离最简单的计算方法。它本身有很多限制，比如欧式距离并没有考虑向量中每个项目的权重，而是直接直接平方相加。

**余弦相似度适用场景**

余弦相似度通过向量的方向计算相似度，通常在推荐系统用户-商品评分、自然语言处理（NLP）等场景中使用。但是它本身也存在一些缺陷。比如余弦相似度在两个向量方向一致时相似度为1，这很显然是不完全正确的，比如上文中[0,2,0,2,0,2]和[0,4,0,4,0,4]这两个item的评分向量虽然方向相同，夹角为0，但相似度为1很显然不正确。

**对比欧式和余弦相似度**

对比欧式距离的计算和夹角余弦的计算结果发现，欧式距离的相似度比较小，而夹角余弦相似度比较大，即夹角余弦更能反映两者之间的变动趋势，两者有很高的变化趋势相似度，而欧式距离较大是因为两者数值有很大的区别，即两者拥有很高的数值差异。

欧氏距离和余弦距离各自有不同的计算方式和衡量特征，因此它们适用于不同的数据分析模型。欧氏距离能够体现个体数值特征的绝对差异，所以更多的用于需要从维度的数值大小中体现差异的分析，如使用用户行为指标分析用户价值的相似度或差异。

余弦距离更多的是从方向上区分差异，而对绝对的数值不敏感，更多的用于使用用户对内容评分来区分兴趣的相似度和差异，同时修正了用户间可能存在的度量标准不统一的问题（因为余弦距离对绝对数值不敏感）。

而在实际的推荐系统应用中，评分矩阵通常是高维且稀疏的，所以计算余弦相似度比欧式距离更合适。

**皮尔森系数使用场景**

皮尔森系数，与夹角余弦类似，但是可以去中心化。比如评分时，有人倾向于打高分，有人倾向于打低分，他们的最后效果在皮尔森中是一样的。

**曼哈顿距离使用场景**

曼哈顿距离，一般在路径规划、地图类中常用，比如A\*搜索算法[[14]](#footnote-14)中使用曼哈顿来作为每一步代价值的一部分（F=G+H, G是从当前点移动到下一个点的距离，H是距离目标点的距离，这个H就可以用曼哈顿距离表示）

在某些情况下，如果单一计算方式无法满足系统要求，可以计算两种或者多种相似度，然后设置不同的权重最终累加。但是，这一方法要是具体情况而定，因为每中相似度数值的量纲可能不一样，需要进行归一化操作，不一定能直接做加法运算。

## 基于评分的相似度计算实现

在推荐系统中，如果有user对item的评分，那就可以就行相似度计算了，其中包括item相似度与user相似度。

|  |  |  |  |
| --- | --- | --- | --- |
| U\I | Item1 | Item2 | Item3 |
| User1 | 4 | 3 | 5 |
| User2 | 2 | 5 | 4 |
| User3 | 2 | 4 | 0 |

根据上述评分矩阵，其中User3关于Item3的评分是未知的，先设置为0。User1的评分向量为[4,3,2], User3的评分向量为[2,5,4],计算User1和User3的余弦相似度为（4\*2+3\*4+5\*0）/() 0.632。User2的评分向量为[2,5,4]，计算User2和User3的余弦相似度为（2\*2+4\*4+4\*0）/() 0.800。从计算结果看出，与User3最相似的用户是User2.

但是，真实场景中用户点击或者购买某一商品时并不会对商品打分，这时可以通过一些方法进行构造式评分。

|  |  |
| --- | --- |
| **行为** | 评**分** |
| 点击 | 2.0 |
| 浏览详情 | 3.0 |
| 收藏 | 3.5 |
| 加购 | 4.0 |
| 下单 | 4.5 |

另外，有些推荐场景并没有用户点击、加购等复杂的操作行为，只有购买行为。这时可以用0和1表示用户的购买情况，0表示未买过，1表示买过。下图就是将购买行为转换成0和1的矩阵。

|  |  |  |  |
| --- | --- | --- | --- |
| U\I | Item1 | Item2 | Item3 |
| User1 | 1 | 0 | 1 |
| User2 | 0 | 1 | 0 |
| User3 | 1 | 0 | 0 |

如果要得到标准的1-5的评分矩阵，需要将用户购买物品的次数转换成评分。数据归一化到[a,b]区间范围的方法：

（1）首先找到样本数据X的最小值Min及最大值Max  
（2）计算系数为：k=（b-a）/(Max-Min)  
（3）得到归一化到[a,b]区间的数据：norY=a+k(X-Min)

用python实现归一化操作为：  
def Normalization(x,a,b):  
 k = (b-a)/float(max(x)-min(x))  
 vec = [a+k\*(float(i)-min(x))/float(max(x)-min(x)) for i in x]  
 return [round(i, 2) for i in vec]

有上述方法转换得到的评分再取小数点后两位得到评分矩阵：

|  |  |  |  |
| --- | --- | --- | --- |
| User | Item | count | Rating |
| User1 | Item1 | 4 | 1.44 |
| User2 | Item2 | 10 | 1.15 |
| User3 | Item3 | 1 | 1.0 |

在MovieLens电影数据ml-20m中，计算所有元素的笛卡尔积，得到所有元素两两计算的相似度。这种常见的计算方式比较原始，也是非常耗时的。比如ratings.csv文件中Item去重后个数为2万多个，用户的维度有13万。如果计算Item的相似度，需要计算4亿次，时间复杂度较高。另外，由于用户user维度更大，两两计算相似度，那总共就要算次相似度，计算效率是个亟需解决的问题。存储方面，计算user相似度时，假设每次计算保存4个Byte，总共需要70G的数据量，空间存储也是问题。

### RDD分布式计算

当前代码中实现的相似度计算是通过笛卡尔积的方式计算RDD中两两匹配的相似度。RDD作为分布式计算数据集，可以很容易的实现多线程、多节点的进行相似度计算，比单机单进程的计算方式要快很多。一般业务场景下item数量要远远大于user数据量。所以，在大数据量下，即时单机可以容忍itemCF的计算时间，userCF则不行。最简单的RDD优化方式，可以通过广播变量的方式提高对比计算的效率。

### 向量化计算

根据余弦相似度的计算公式：

可以上述公式转换成向量形式，如下：

为什么要进行向量化。向量化计算是机器学习领域常用的方法，也就是把循环遍历转换成向量计算，由此提高计算效率。在前面章节中介绍了Spark MLlib中的向量数据类型，我们用分布式矩阵CoordinateMatrix实现一个真实场景的Item相似度计算案例。

通过CoordinateMatrix进行向量相似度计算scala代码：

import org.apache.spark.mllib.linalg.distributed.{CoordinateMatrix, MatrixEntry}  
val df = spark.createDataFrame(Seq(  
 (0, 0, 1.0),  
 (1, 0, 1.0),  
 (2, 0, 1.0),  
 (3, 0, 1.0),  
 (0, 1, 2.0),  
 (1, 1, 2.0),  
 (2, 1, 1.0),  
 (3, 1, 1.0),  
 (0, 2, 3.0),  
 (1, 2, 3.0),  
 (2, 2, 3.0),  
 (0, 3, 1.0),  
 (1, 3, 1.0),  
 (3, 3, 4.0)  
))  
  
val matrix = new CoordinateMatrix(df.map(row =>  
 MatrixEntry(row.getAs[Integer](0).toLong,  
 row.getAs[Integer](1).toLong,  
 row.getAs[Double](2))).toJavaRDD)  
// 计算similarity  
val result = matrix.toRowMatrix().columnSimilarities()  
// 打印相似度结果  
result.entries.  
 map(x => (x.i, x.j, x.value)).collect().foreach(println)

结果输出：

(2,3,0.2721655269759087)

(1,2,0.9128709291752768)

(1,3,0.596284793999944)

(0,3,0.7071067811865476)

(0,1,0.9486832980505139)

(0,2,0.8660254037844386)

按照以上案例，把向量化的相似度计算方法用到评分矩阵的计算中。对应的scala代码如下：

val matrix = new CoordinateMatrix(df.map(row =>  
 MatrixEntry(row.getAs[Integer](0).toLong,  
 row.getAs[Integer](1).toLong,  
 row.getAs[Double](2))).toJavaRDD)  
// 调用sim方法  
val x = matrix.toRowMatrix().columnSimilarities()  
// 得到相似度结果  
x.entries.collect().foreach(println)

//利用Spark分布式矩阵计算Item相似度

def itemVectorSimilarity(spark: SparkSession,  
 trainData: Dataset[UserItem],  
 topN: Int) = {  
 //计算相似度矩阵  
 def standardCosine(matrix: CoordinateMatrix): RDD[MatrixEntry] = {  
 val similarity = matrix.toIndexedRowMatrix().columnSimilarities()  
 val sim = similarity.entries  
 sim  
 }  
  
 val sim = standardCosine(parseToMatrix(trainData)).map {  
 case MatrixEntry(user1, user2, sim) => (user1, (user2, sim))

}.groupByKey().

flatMapValues {x =>  
 val sim\_users = x.toList.sortBy(-\_.\_2).take(topN)  
 sim\_users  
 }.map(x => (x.\_1, x.\_2.\_1, x.\_2.\_2))  
}

### 近似最近邻方法ANN

ANN并非人工神经网络，而是近似邻居算法（Approximate Nearest Neighbor）的缩写。根据之前的方法，计算笛卡尔积属于蛮力（brute-force）搜索的方式，也就是在全空间进行搜索，为了加快查找的速度，几乎所有的ANN方法都是通过对全空间分割，将其分割成很多小的子空间，在搜索的时候，通过某种方式，快速锁定在某一（几）子空间，然后在该（几个）子空间里做遍历。可以看到，正是因为缩减了遍历的空间大小范围，从而使得ANN能够处理大规模数据的索引[[15]](#footnote-15)。关于ANN的算法实现有很多，比较典型的实现有Annoy、nmslib和SPTAG等等。几乎所有的ANN算法都是对于全空间的划分，大多数使用的是树模型。

1. LSH

局部敏感哈希（Locality-Sensitive Hashing）简称LSH是一种用于求解高维空间中的近似近邻搜索的方法。LSH保证了高维空间相近的点映射到低维空间相近的概率很高，所以这也是一个降维的过程。

1. K-Means Tree

K-Means Tree实际就是对数据做了多层K-means，每一层到当前的划分“叶子节点”包含样本数都小于1个。

1. K-D Tree

是对数据点在k维空间中划分的一种数据结构。k-d tree实际上是一种二叉树。

#### Annoy介绍：

Annoy是近似最近邻算法实现的一个C++ 库，它是（Approximate Nearest Neighbors Oh Yeah）的缩写，其本身带有Python接口。用于搜索空间中给定查询项目的的近似项目。它也创建了大型只读文件的数据结构，这些数据结构被mmap[[16]](#footnote-16)到内存中，以便许多进程可以共享相同的数据。

安装

安装很简单，只需通过pip安装：

pip install—user annopy

这行命令会从PyPI上拉下Annoy最新版本。

对于C++版本，只需从github中克隆项目，并插入声明#include “annoylib.h“即可。

**背景**

还有一些其他库可以实现最近的邻域搜索。Annoy几乎是执行速度最快的那一个（下面会做一些性能对比），但实际上还有另一个使得Annoy与众不同的功能：它能够使用静态文件作为索引。这意味着您可以跨进程共享索引。Annoy还会将创建索引与加载索引分离，以便您可以将索引作为文件传递并快速映射到内存中。Annoy的另一个好处是，它尝试最小化内存占用，因此索引非常小。

这个功能相当好用，如果要查找最近邻，并且有许多CPU，则只需生成一次索引。您还可以传递和分发静态文件，以用于生产环境、Hadoop作业等。任何进程都能够将索引加载（mmap）到内存中，并能够立即进行查找。

该项目在Spotify[[17]](#footnote-17)上使用并进行音乐推荐。在运行矩阵分解算法后，每个用户/项目都可以在f维空间中表示为向量。Annoy帮助系统搜索相似的user/item，这是一个超过百万维度的空间，因此内存使用是首要问题。

**功能摘要：**

1. 欧几里德距离，曼哈顿距离，余弦距离，汉明距离(Hamming Distance)或点（内）积距离。

2. 余弦距离相当于归一化向量的欧几里德距离= sqrt（2-2 \* cos（u，v））

3. 如果维度很小（比如小于<100维），效果会更好，但即使是1,000维也能表现出色

4. 内存使用量小

5. 允许您在多个进程之间共享内存

6. 索引创建与查找分开（特别是在创建树后，您无法添加更多项目）

7. Python支持2.7,3.6和3.7。

8. 如果数据量很大，可以不用放到内存中，在磁盘上构建索引再使用

利用python调用计算item相似度的案例：

from annoy import AnnoyIndex  
class AnnoyMovie:  
 def \_\_init\_\_(self, dataFile):  
 self.datafile = dataFile  
 self.data = self.loadData()  
 self.ann, self.userNum = self.addItem()  
  
 def loadData(self):  
 print(“Load data...”)  
 data = []  
 for line in open(self.datafile):  
 itemid, vector = line.strip(“\n”).split(“,”)  
 data.append((itemid, vector))  
 return data  
  
 def addItem(self):  
 userNum = len(self.data.pop()[1].split(“ “))  
 print(“user num”, userNum)  
 ann = AnnoyIndex(userNum, ‘angular’)  
 for itemId, vector in self.data:  
 ann.add\_item(int(itemId), map(float,vector.split(“ “)))  
 ann.build(1000)  
 return ann, userNum  
  
 def getNear(self, index, itemId, simNum):  
 print(index.get\_nns\_by\_item(itemId, simNum))  
  
 def save(self):  
 print(“保存索引文件...”)  
 self.ann.save(‘movie.ann’)  
  
 def load(self):  
 print(“加载索引文件...”)  
 u = AnnoyIndex(self.userNum, ‘angular’)  
 u.load(‘movie.ann’)  
 return u

注意：Annoy只接受整数作为Item的标识符。请注意，它将为max（id）+1个Item分配内存，因为它假设您的Item编号为0 ... n-1。

**部分python API接口：**

AnnoyIndex（f，metric）

返回一个新的读写索引并存储f维度的向量。相似度计算可以是“angular”（类似于cosine距离），“欧几里德”，“曼哈顿”，“汉明”或“点积”。

a.add\_item（i，v）

用向量v添加项i（任何非负整数）。注意它将为max（i）+1项分配内存。

a.build（n\_trees）

构建一个n\_trees树的森林。查询时，更多树提供更高的精度。调用build后，不能再添加任何项目。

a.save（fn，prefault = False）将索引保存到磁盘并加载它（参见下一个函数）。保存后，不能再添加任何项目。

a.load（fn，prefault = False）

从磁盘加载索引。如果prefault设置为True，它将预读取整个文件到内存中。默认值为False。

a.unload（）为卸载。

a.get\_nns\_by\_item（i，n，search\_k = -1，include\_distances = False）

返回n个最接近的项目。在查询期间，它将检查最多search\_k节点，如果没有提供，则默认为n\_trees \* n。 search\_k为您提供更好的准确性和速度之间的运行时权衡。如果将include\_distances设置为True，它将返回一个包含两个列表的2元素元组：第二个元素包含所有相应的距离。

get\_nns\_by\_vector（v，n，search\_k = -1，include\_distances = False）

相同但是按向量v查询。 a.get\_item\_vector（i）返回先前添加的项目i的向量。 a.get\_distance（i，j）返回项i和j之间的距离。

#### hnswlib介绍[[18]](#footnote-18)

在说hnslib之前先说nmslib。非度量空间库（nmslib）是一个高效的跨平台的相似性搜索库，也是用于评估相似性搜索方法的工具包。其核心库没有任何第三方依赖。Nmslib是一个可扩展的库，这意味着可以添加新的搜索方法和距离函数。Nmslib可以直接在C++和python中调用。另外，还可以构建一个查询服务器，可以用Java调用。

快速近似邻居搜索，Hnswlib与nmslib来自于同一作者，而hnswlib更快，占用内存更小。

明显优势：

1. 体积小，调用方便，C++11开发且没有其他任何依赖

2. 接口包含C++，python和R

3. 支持新增索引，同时支持元素删除

4. 可以自定义距离计算

5. 与nmslib相比，明显减少了内存占用，同时缩短了构建时间

安装：

pip install hnswlib

利用python调用hnswlib实现item相似度计算：

import hnswlib  
import numpy as np  
import sys  
  
class HnmslibMovie:  
 def \_\_init\_\_(self, dataFile):  
 self.datafile = dataFile  
 self.data, self.data\_labels = self.loadData()  
 self.ann = self.addItem()  
  
 #导入评分矩阵文件  
 def loadData(self):  
 print(“Load data...”)  
 data = []  
 data\_labels = []  
 for line in open(self.datafile):  
 index, vector = line.strip(“\n”).split(“,”)  
 rating\_list = vector.split(“ “)  
 data\_labels.append(index)  
 data.append(rating\_list)  
 data = np.float32(data)  
 data\_labels = np.int32(data\_labels)  
 return data, data\_labels  
  
 def addItem(self):  
 self.dim = len(self.data[0])  
 self.num\_elements = len(self.data\_labels)  
 print(“dimension num”, self.dim)  
 #声明索引  
 p = hnswlib.Index(space=’cosine’, dim= self.dim) #space设置为l2, cosine或ip  
 #初始化索引，应事先知道最大数量的元素  
 p.init\_index(max\_elements= self.num\_elements, ef\_construction=200, M=16)  
 #插入元素（可以被调用多次）  
 p.add\_items(self.data, self.data\_labels)  
 #设置ef，控制召回  
 p.set\_ef(50) # ef设置必须要大于k  
 return p  
  
 def getNear(self, p, index, simNum):  
 #查询近邻，k=最近元素的个数（返回两个numpy数组）  
 labels, distances = p.knn\_query(self.data[np.where(self.data\_labels==index)[0][0]], k=simNum)  
 print(“label:”,labels, “distance: “, distances)  
 return labels, distances  
  
 def save(self):  
 print(“保存索引文件...”)  
 self.ann.save\_index(‘movie.ann’)  
  
 def load(self):  
 print(“加载索引文件...”)  
 p = hnswlib.Index(space=’cosine’, dim=self.dim)  
 p.load\_index(‘movie.ann’, max\_elements = self.num\_elements)  
 return p  
  
 def localQuery(self):  
 self.save()  
 p = self.load()  
 # 制定itemid和 相似度的个数simNum  
 self.getNear(p, 1, 10)

#### SPTAG介绍

微软开源了搜索引擎Bing的搜索算法，称为空间分区式的树和图（SPTAG，Space Partition Tree And Graph），也是一种分布式近似最近邻搜索（ANN）库。可以为大规模向量搜索场景提供高质量向量索引构建。这里“树”为KD-Tree和K-Means树，“图”是相对邻域图（SPTAG-BKT）。SPTAG-KDT在指数构建成本方面很有优势，而SPTAG-BKT在超高维数据中的搜索精度方面很有优势。

运用在推荐系统中，假设用户-物品的评分表示为向量，通过L2距离（欧式距离）或余弦距离来比较向量。而查询向量返回的向量是与查询向量具有最小L2距离或余弦距离的向量。

**SPTAG工作原理：**

搜索算法一般分为索引（index）查询（search）两个部分。SPTAG的索引是基于kd-tree。其在于一维空间上的情况就是“平衡二叉树”，在高维空间上kd-tree会用第k维的大小来决定一个元素应该插入左子树还是右子树，同时为保持tree的平衡，剩余未进入tree的元素除第k维外方差最小。SPTAG正是以此来加速算法的速度。

另外，k-means是一种自动聚类的方法，算法先随机指定选取K个点做为初始聚集的簇心,分别计算每个样本点到K个簇核心的余弦距离，找到距离最近的核心点，将它归属到对应的簇，所有点都归属到簇之后， M个点就分为了K个簇。之后重新计算每个簇的重心，将其定为新的“核心”，重复上述步骤直到新核心不再改变为止或者改变距离达到一定值后中止。那么最终的K个簇就是最终的聚类结果。

结合kd-tree和k-means这两种算法，SPTAG允许用户在几毫秒内搜索数十亿条信息。

安装SPTAG

安装系统要求：

gcc >= 5.0

swig >=3.0

cmake>= 3.12.0

boost >= 1.67.0

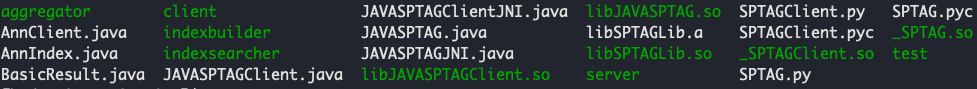
如果你的系统gcc版本过低，先升级gcc然后再升级cmake

从github地址[[19]](#footnote-19)下载源码，进入SPTAG-master目录，执行命令：

mkdir build

cd build && cmake .. && make

安装完成之后SPTAG-master目录下回生成一个Release/文件夹，其中包含一些可执行文件、Java和Python的API文件。



本节只介绍python的调用方式，其他方式可以docs文档中查看。

先用MovieLens的数据集ml-1m上进行测试，通过SPTAG的python API进行近似最近邻找找。找出与电影《玩具总动员》[1,Toy Story (1995),Animation|Children’s|Comedy]最相似的Top10个其他电影。

第一步骤：按照格式生成原始数据

假设向量的维度为3格式如下：

<metadata1>\t<v11>|<v12>|<v13>|

<metadata2>\t<v21>|<v22>|<v23>|

...

其中每行表示一个向量，其元数据及其值由制表符分隔。向量的每个维度都由 |或使用—分隔。

第二步骤：按照格式生成查询文件

查询文件的格式与上面相同。

第三步骤：构建索引

第四步骤：查询索引

下面是python代码构建索引和查询索引。

import sys  
sys.path.append(‘/myfile/SPTAG-master/Release/’)  
import SPTAG  
import numpy as np  
  
class SPTAG\_Movie:  
 def \_\_init\_\_(self, dataFile):  
 self.datafile = dataFile  
 self.data, self.data\_labels = self.loadData()  
 # self.testBuild(“KDT”, “cosine”, self.data, ‘testindices’)  
 # self.testSearch(‘testindices’, q, k)  
  
 def loadData(self):  
 print(“Load data...”)  
 data = []  
 data\_labels = []  
 for line in open(self.datafile):  
 index, vector = line.strip(“\n”).split(“,”)  
 rating\_list = vector.split(“ “)  
 data\_labels.append(index)  
 data.append(rating\_list)  
 data = np.float32(data)  
 data\_labels = np.int32(data\_labels)  
 return data, data\_labels  
  
 def testBuild(self, algo, distmethod, x, out):  
 i = SPTAG.AnnIndex(algo, ‘Float’, x.shape[1])  
 i.SetBuildParam(“NumberOfThreads”, ‘4’)  
 i.SetBuildParam(“DistCalcMethod”, distmethod)  
 ret = i.Build(x, x.shape[0])  
 i.Save(out)  
  
 def testSearch(self, index, q, k):  
 j = SPTAG.AnnIndex.Load(index)  
 for t in range(q.shape[0]):  
 result = j.Search(q[t], k)  
 result = [x+1 for x in result[0]]  
 print (self.idToName(result)) #索引ids+1=movieID  
 print (result[1]) # distances  
  
 def idToName(self, list):  
 print(“读取id-name文件...”)  
 data = {}  
 with open(“../ml-1m/movies.csv”) as f:  
 # for line in open(“../ml-1m/movies.csv”):  
 next(f)  
 for line in f:  
 sp = line.strip(“\n”).split(“,”)  
 data[int(sp[0])] = sp[1]+”,”+sp[2]  
 result = [data.get(x,x) for x in list]  
 for i in result:  
 print(i)

if \_\_name\_\_ == ‘\_\_main\_\_’:  
 # ann = SPTAG\_Movie(“/xiaoi/josh/ann/ml\_1m\_vector.txt”)  
 # ann.testBuild(“KDT”, “cosine”, ann.data, ‘testindices’)  
 ann = SPTAG\_Movie(“/xiaoi/josh/ann/query\_input.txt”)  
 ann.testSearch(‘testindices’, ann.data[0], 10)

输出结果得到与电影玩具总动员的最相似的十个电影：

[1, 2899, 1174, 575, 2163, 1179, 34, 1107, 347, 2375]

转为Movie-name:

Toy Story (1995),Animation|Children’s|Comedy

Gulliver’s Travels (1939),Adventure|Animation|Children’s

Grosse Fatigue (1994),Comedy

Little Rascals, The (1994)

Attack of the Killer Tomatoes! (1980),Comedy|Horror

Grifters, The (1990)

Babe (1995),Children’s|Comedy|Drama

Loser (1991),Comedy

Bitter Moon (1992),Drama

Money Pit, The (1986)

#### 矩阵分解MF与ANN结合

在实际推荐系统中，高维度的超大稀疏矩阵在单机情况下很难构建索引进行ANN最近邻搜索。主要是因为内存资源不够，即使通过磁盘存储来缓解内存压力，整体构建索引的时间也是非常漫长的。在建构索引之前，有没有一种方法能够降低数据的维度又不会损失太多的信息呢？数据降维就是一种有效的方法。这里介绍一种基于矩阵分解（MF）的方法。

libMF是一个关于推荐系统的矩阵分解库。它会利用用户和物品的评分矩阵训练出一个模型，然后就可以通过该模型预测评分了。在这个过程中会生成两个矩阵，分别是用户矩阵P和物品矩阵Q。每个矩阵的维度是可以指定的。

libMF在进行矩阵分解之后会生成一个模型文件，前几行如下：

f 0

m 88997

n 33183

k 1000

b 1.41962

这几行主要记录用户和物品的维度信息等等，读取矩阵的时候可以跳过。

下面读取LibMF矩阵分解之后生成的Item矩阵，然后再用Annoy构架索引：

def loadMF(self):  
 print(“Load mf data...”)  
 data = []  
 input\_file = open(self.datafile)  
 for line in islice(input\_file, 5, None):  
 list = line.strip().split(“ “)  
 itemId = list[0]  
 flag = list[1]  
 vector = “ “.join(list[2:])  
 if(itemId[0] == “q” and flag == “T”):  
 data.append((itemId[1:], vector))  
 return data

#### SPTAG高级操作

1. **增量添加和删除**

实时增加和删除向量是SPTAG独有的一个功能，这一功能可以利用在推荐系统的实时推荐中。通过增加向量，系统可以实时更新用户的最新行为，也可以增加新注册用户的特征。

下面是SPTAG增加和删除向量的操作：

def addItem(self, index, x, out, algo, distmethod):  
 if index != None:  
 i = SPTAG.AnnIndex.Load(index)  
 else:  
 i = SPTAG.AnnIndex(algo, ‘Float’, x.shape[1])  
 i.SetBuildParam(“NumberOfThreads”, ‘4’)  
 i.SetBuildParam(“DistCalcMethod”, distmethod)  
 if i.Add(x, x.shape[0]):  
 i.Save(out)  
  
def deleteItem(self, index, x, out):  
 i = SPTAG.AnnIndex.Load(index)  
 ret = i.Delete(x, x.shape[0])  
 print (ret)  
 i.Save(out)

1. **实时API访问查询**

案例：电影SPTAG实时查询最近邻服务：

步骤1：配置servic.ini

[Service]

ListenAddr=122.226.240.159

ListenPort=8001

ThreadNumber=8

SocketThreadNumber=8

[QueryConfig]

DefaultMaxResultNumber=100

DefaultSeparator=|

[Index]

List=KDT //索引的类型

[Index\_KDT]

IndexFolder=testindices //索引文件

步骤2：启动server服务

启动server，正常情况下会打印导入索引的日志，如果没有则是配置文件有问题。

/xiaoi/josh/SPTAG-master/Release/server -m socket -c ./service.ini

步骤3：调用SPTAG\_Client读取本地查询文件搜索最近邻

def testSPTAGClient(self, q, k):  
 index = SPTAGClient.AnnClient(‘122.226.240.159’, ‘8001’)  
 while not index.IsConnected():  
 time.sleep(1)  
 index.SetTimeoutMilliseconds(18000)  
 for t in range(q.shape[0]):  
 result = index.Search(q[t], k, ‘Float’, False)  
 result = [x+1 for x in result[0]]  
 print (self.idToName(result)) #索引ids+1=movieID  
 print (result[1]) # dis

以下是通过8001端口查询电影《Toy Story》Top10最近邻的输出结果：

Toy Story (1995),Animation|Children’s|Comedy

Gulliver’s Travels (1939),Adventure|Animation|Children’s

Grosse Fatigue (1994),Comedy

Little Rascals, The (1994)

Attack of the Killer Tomatoes! (1980),Comedy|Horror

Grifters, The (1990)

Babe (1995),Children’s|Comedy|Drama

Loser (1991),Comedy

Bitter Moon (1992),Drama

Money Pit, The (1986)

以下是增加和删除向量的日志：

Add 1 vectors

Save Vector To testindices/vectors.bin

Save Vector (3708, 6041) Finish!

Save KDT to testindices/tree.bin

Save KDT (1,3706) Finish!

…

Load Vector From testindices/vectors.bin

Load Vector (3708, 6041) Finish!

Load KDT From testindices/tree.bin

Load KDT (1,3706) Finish!

Load Graph From testindices/graph.bin

Load Graph (3708, 64) Finish!

Load DeleteID From testindices/deletes.bin

Load DeleteID (0) Finish!

True

Save Vector To testindices/vectors.bin

Save Vector (3708, 6041) Finish!

Save KDT to testindices/tree.bin

Save KDT (1,3706) Finish!

Save Graph To testindices/graph.bin

Save Graph (3708, 64) Finish!

Save DeleteID To testindices/deletes.bin

Save DeleteID (1000) Finish!

#### Spark-MinHashLSH

局部敏感哈希（Locality Sensitive hash，LSH），主要用来解决海量数据的相似性检索。在Spark MLlib中LSH使用一系列函数将数据点哈希到桶中，使得彼此接近的数据点在相同的桶中具有高概率，而数据点是远离彼此很可能在不同的桶中。Spark中LSH目前只支持欧式距离与Jaccard距离。

Spark实现LSH具体对象为：BucketedRandomProjectionLSH和MinHash。

Spark-LSH实现Item相似度计算，Scala代码如下：

import org.apache.spark.ml.linalg.Vectors

import org.apache.spark.ml.linalg.{SparseVector, Vectors}  
import org.apache.spark.ml.feature.{BucketedRandomProjectionLSH, MinHashLSH, MinHashLSHModel}  
  
val user\_count = data\_rating.map(\_.\_1.toInt).max  
val data\_rating\_matrix = data\_rating.map (x => (x.\_2.toInt, (x.\_1.toInt, x.\_3))).  
 groupByKey().mapValues { x =>  
 val features = x.toList.map {  
 case (feature\_index, value) =>  
 (feature\_index, value)  
 }.sortBy(\_.\_1)  
 Vectors.sparse(user\_count - 1, features)  
}.sortByKey()  
val data\_rating\_df = spark.createDataFrame(data\_rating\_matrix).toDF(“id”, “features”)

val mh = new MinHashLSH().  
 setNumHashTables(5).  
 setInputCol(“features”).  
 setOutputCol(“hashes”)  
  
val itemModel = mh.fit(data\_rating\_df)  
val transformed\_data\_rating = itemModel.transform(data\_rating\_df)  
transformed\_data\_rating.show(false)

val itemSimRDD = data\_rating\_matrix.collect().  
 map(x=> (x.\_1, itemModel.approxNearestNeighbors(data\_rating\_df, x.\_2, 10)))

itemSimRDD

LSH的具体参数：

Tables： 该参数越大，可以降低误报率（false negative rate）；参数越小，可以提高运行的效率。

### 实验对比与最佳方式

对比上述几种相似度计算方式，由于user比item相似度计算更加耗时，且userID无法进行抽样检查。这里选择item进行相似度对比。

分别用单机和集群两种资源配置方式进行测试。

单机配置：CPU-24cores, Memory-64G

Spark-Local限制资源：CPU-20cores，memory-30G

集群配置：Master+8worders, 单个worker：CPU-24cores, Memory-64G

Spark-Cluster限制资源：8workers，CPU-160cores，Executor-19G

**Annoy运行命令：**

140机器： /xiaoi/josh/recommend/ann

运行ml\_1m数据：

time python ann\_movie\_vector.py

运行CRM数据:

time python ann\_crm\_vector.py

hnswlib运行命令：

140机器： /xiaoi/josh/recommend/ann

运行ml\_1m数据：

time python hnswlib\_movie.py

运行CRM数据:

time python hnswlib\_crm.py

**SPTAG运行命令：**

159机器：/xiaoi/josh/ann/movie

构建索引：

time python sptagMovie.py

查询：

1. 修改读取文件query\_input.txt
2. 执行

python sptag\_load.py

注：

1.“/”表示资源不支持该运算，或者是无法运行。

2.“\_decom”后缀表示数据降维后再进行计算，sklearn - PCA(n\_componets取0.99)

3. 运行时间再具体测试时会稍微上下浮动。

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Data | 计算方式 | 内存占用 | 构建索引 | 加载并查询一次(Top10) | 加载并查询1千次(Top10) | 加载并查询1万次(Top10) |
| ml-1m | Annoy | 0.2G | 6.104s | 6.15s | 19.631s | / |
| hnswlib | 1.1G | 10.211s | 10.809s | 64.57s | / |
| SPTAG | 0.2G | 15m23 | 0.4s | 6.909s | / |
| CRM | Annoy | 24G | 12m12s | 13m25s | 58m25s | >2h |
| hnswlib | >60G | / | / | / | / |
| hnswlib\_decom | >60G | / | / | / | / |
| SPTAG | >60G | / | / | / | / |

快速ANN算法工具库实验小结：

**内存方面：**

对比Annoy、hnswlib和SPTAG三者在不同数据量下的内存和时间消耗.我们发现，Annoy并不会因为数据量过大而把内存全部占尽，内存最高占用24G，如果不够则把部分数据存储在磁盘当中。相反，hnswlib在大数据量下因为内存不够用，程序被迫中止。STPTAG单机试验下占用内存超过最高上限，报错MemoryError。

**速度方面：**

由于Hnswlib和SPTAG单机状态下，无法实现CRM数据（维度20000\*80000）的计算。只能通过数据ml\_1m(维度3700\*6000)小数据集下进行对比。

**效果方面：**

Annoy和Hnswlib将参数都选择cosine余弦距离。抽取了几个例子，两者相似度结果保持一致。

但是SPTAG的相似度计算结果与Annoy和Hnswlib大范围不同。比如：

同时计算Index1，玩具总动员的相似Top10：

**Annoy的结果:**

[1, 3114, 1265, 588, 2355, 1270, 34, 1196, 1580, 356]

转为Movie-name:

Toy Story (1995),Animation|Children’s|Comedy

Toy Story 2 (1999),Animation|Children’s|Comedy

Groundhog Day (1993),Comedy|Romance

Aladdin (1992),Animation|Children’s|Comedy|Musical

Bug’s Life, A (1998)

Back to the Future (1985),Comedy|Sci-Fi

Babe (1995),Children’s|Comedy|Drama

Star Wars: Episode V - The Empire Strikes Back (1980),Action|Adventure|Drama|Sci-Fi|War

Men in Black (1997),Action|Adventure|Comedy|Sci-Fi

Forrest Gump (1994),Comedy|Romance|War

**Hnswlib的结果：**

[1, 3114, 1265, 588, 2355, 1270, 34, 1196, 1580, 356]

**SPTAG结果：**

[1, 2899, 1174, 575, 2163, 1179, 34, 1107, 347, 2375]

转为Movie-name:

Toy Story (1995),Animation|Children’s|Comedy

Gulliver’s Travels (1939),Adventure|Animation|Children’s

Grosse Fatigue (1994),Comedy

Little Rascals, The (1994)

Attack of the Killer Tomatoes! (1980),Comedy|Horror

Grifters, The (1990)

Babe (1995),Children’s|Comedy|Drama

Loser (1991),Comedy

Bitter Moon (1992),Drama

Money Pit, The (1986)

可以看出Annoy的结果与Hnswlib高度一致，而SPTAP的结果与前两者大相径庭。

**Spark Item相似度计算方法时间对比：**

|  |  |  |  |
| --- | --- | --- | --- |
| Data | 计算方式 | 单机Local | 集群Cluster |
| CRM | RDD | 60min | 9min |
| Spark-Vectorization | 25.39min | 2.6min |
| Spark-LSH | 14.11min | 3.8min |

当前最佳方案： ANN + TopN。考虑到实际场景中需要将所有Item的相似度TopN进行召回，以便于下一步离线排序使用。但是，不再需要计算所有Item的笛卡尔积，然后再排序选取TopN。只要查找单个Item的最近邻相似TopN即可。时间复杂度大大降低。由此，真实推荐中大规模高维度的相似度计算问题，可以通过近似最近邻方式快速找到相似（接近）的用户和物品。

影响相似度计算时间的因素：

1. User和Item的数量
2. 用户产生的行为数量，这直接影响数据集的稀疏度
3. 硬件资源，理论上将节点越多，计算速度越快

**矩阵分解MF与ANN结合**

实验： 将CRM的评分数据集进行MF矩阵分解，再用Annoy和Sptag构建索引，查询Item近似最近邻物品

目录和文件：/xiaoi/josh/recommend/ann

数据： CRM

机器：140，24cores, 64G

相关命令：

time /xiaoi/recommend/project/install\_package/libmf-2.01/mf-train crm\_rating.txt crm\_libmf.model

python ann\_crm\_libmf.py 6760 //传参指定查询的Item ID

图片包含 屏幕截图

描述已自动生成

接上图：

图片包含 文字, 屏幕截图

描述已自动生成

根据LibMF矩阵分解之后在进行ANN搜索的实验结果，可以看出矩阵分解的误差越小，就需要增加迭代次数和矩阵维度（隐因子数量）。随之而然的就是训练时间的增加。从简单的抽样结果看，矩阵分解之后的向量最近邻结果，很难与原始余弦相似度的结果一致。

### 常见问题与解决方法

**稀疏向量的压缩**

实际生产环境中，生成的用户向量都是非常稀疏的，构成的矩阵也是非常稀疏的，直白来说就是很多值都是0，有一些存储稀疏矩阵的格式。比如CSR或者COO。

CSR：CSR是一个整体编码方式，由三部分构成，数值、列号和行偏移。

COO：COO每个元素用一个三元组表示（行号，列号，数值），只存储有值的元素，缺失值不存储。

这些存储格式，在常见的框架中都已经实现，比如Python-scipy和Spark-MLlib。

**矩阵降维**

首先，相似度计算本身并不需要计算全量数据，比如超过3个月的行为数据就不应该作为计算数据，因为用户的行为习惯是变化的，太老的数据已经不能反映客户当前偏好和习惯。

协同过滤在实际场景中itemCF效果会好于UserCF，一般情况下只需要计算Item相似度即可。User相似计算一般会在计算用户特征相似时使用，找出新客户最相似的老客户就可以推荐老客户购买过的商品，这一方法可以解决用户冷启动的问题。

## 基于文本的相似度

### 文本相似的度量方式

文本相似度计算方法：

图片包含 屏幕截图

描述已自动生成

总的来说，文本相似度量方法可以分为两大类:[[20]](#footnote-20)

String Based，即基于待比较的文本本身中的信息，该类方法评估的是”词法“上的相似性，或说朴素的相似性

Corpus Based，即基于一个较大的文本集合中的信息，该类方法评估的是“语义”上的相似性

#### String Based Methods

要比较两个文本的相似性，比较直观的方法是逐个字符比对，看看有多少个字符是一致的—String Based的方法就是从这种思路中发展出来的。其中以字符为单位去比较的方法被统称为Character Based，而以词为单位的则被称为Term Based。

Character Based的方法普遍用来进行较短文本、小规模的比较，会注意文本中各字符的顺序和位置；Term Based的方法则更适合用来进行较长文本或大规模的比较，而且通常会抛弃文本中各个单元(通常是词)之间的顺序、位置信息—一般的做法是用文本中的词组成的向量来表示文本，也就是所谓的“向量空间模型(Vecter Space Model, VSM)”。

**Character Based Methods**

最长公共子序列(Longest Common Subsequence, LCS)

LCS方法通过计算出两个字符串/序列之间的最长公共子序列，并使用这个子序列的长度来反映两个字符串/序列之间的相似程度。

编辑距离(Leveinshtein Distance)

在两个字符串的比较中应用编辑距离时，通常会有一者作为”标准字符串“，另一者则作为”可能错误“的字符串，通过将后者变换成前者所进行的字符的 插入 、 删除 或 替换 操作次数作为衡量两者差异程度的指标。

扩展的编辑距离(Damerau-Levenshtein Distance)

扩展的编辑距离在思想上与编辑距离一样，只是除插入、删除和替换操作外，还支持 相邻字符的交换 这样一个操作，增加这个操作的考虑是人们在计算机上输入文档时的错误情况中，因为快速敲击而前后两个字符的顺序被输错的情况很常见。

Needleman-Wunsch Similarity

该方法被广泛运用于生物信息学中的序列比对，如氨基酸序列比对、核苷酸序列比对等。其基本思路与编辑距离相近，但在编辑距离中，三种不同的错误情况是平等的，而在生物信息学中，序列中的单元缺失情况比错误(位置匹配但内容不同)情况更不能容忍，因此在4. Needleman-Wunsch方法中，插入错误和删除错误会被赋予较高的惩罚分数。

Smith-Waterman Similarity

Smith-Waterman方法用于生物信息学中的序列比对，但与Needleman-Wunsch方法不一样，它是一个 局部最优比对 方法，简单来说，它的目的是找出两个序列之间 连续且相同 的子序列。

5. Jaro Similarity和Jaro-Winkler Similarity

Jaro方法和Jaro-Winkler方法考虑两个字符串之间相同字符的顺序位置和个数，只适用于像人名这样的较短字符串之间的比较。其中Jaro-Winkler方法是对Jaro方法的改进，而Jaro方法现在已经不常用。

6. Hamming Distance

Hamming距离用于 长度相同 的序列之间的比较，思想非常简单，就是逐位比较得到的不同次数。Hamming距离被广泛应用于信息学。

**Term Based Methods**

Term Based方法中的term不一定是词，也可以是关键词、短语，当选定”词“作为term时，可以用Bag-of-Words Model(BOW) 来更精确地描述此时的模型。为了表述上的方便，本系列文章将此处的term视为词。

1. Cosine Similarity

余弦相似度建立在用向量表示文档的前提上。两个文档的向量，同一个维度应该是表示的同一个词，而每一个维度的值，一般是用”词频-逆文档频率(Term Frequency-Inverse 2. Document Frequency, TF-IDF)“ 来表示。

建立向量表示后，通过计算两个向量之间的夹角大小来衡量两个文档之间的近似程度，由于夹角的余弦值的计算方便，而且天然地处在0和1之间，故一般是用夹角的余弦值而不是夹角的大小来作为度量。

1. Dice’s Coefficient和Jaccard Similarity

Dice系数和Jaccard相似性起初被用于生态学上，作为一种判断物种间相似性的方法。在生态学上，要比较两个物种间相似程度时，通常会对该物种的特性进行采样，最后得到各自的特性集合，而Dice系数和Jaccard相似性都是通过比较两者之间的 共有特性 占比来度量相似性的，因此这两种方法都不是很关心每个 “Term” 的具体量，只是关心有没有某个 “Term”。

#### Corpus Based Methods

Corpus Based方法通过对大量文档的统计分析得到语义上的相似，这里 “大量文档” 就是Corpus即语料了。

**Language Model**

常见的语言模型(Language Model, LM)都基于马尔可夫假设(Markov Assumption)，即认为语言中每个词只与其前面长度为N-1个词有关—这N-1个词其实就构成了该词的上文，同时由于每个词都会成为其他词的上文，下文信息也会得到表现。也就是说，语言模型统计的其实是语言中每个词的一定程度的上下文情况。基于语言模型的这个特点，以及下面这样一个假设:

**具有相同(或相近)上下文的词，其语义是相近的。**

就可以用语言模型来进行文本的相似度量了。具体一点，有两种方式:

通过语言模型来发现同义词、近义词，来弥补Term Based Methods的缺陷；

扩大 “term” 的表示范围，比如说按照词组来进行统计，甚至按照句子来进行统计，那么就可以反映词组之间的相似性、句子之间的相似性了。

**Topic Model**

主题模型与向量空间模型有共同的一点是也是基于BOW模型的，也就是说，并不像语言模型一样考虑词与词之间的顺序、位置，而只是通过词与词的共现(Co-occurrence)来反映词与词之间的相似性。

主题模型是一种无监督的方法，与聚类方法在外在表现上会有一定的相似性，但它们各自的内在原理是相差较大的。

未完待续。。。

### 文本相似度计算

RDD计算杰卡尔德相似度度

编辑距离

/\*\*  
 \* 计算编辑距离  
 \* @param str1待计算的字符串1  
 \* @param str2待计算的字符串2  
 \*/  
def calEditSimilar(str1 : String, str2 : String) : Double = {  
 val factor : Double =  
 if (str1.length >= 8 || str2.length >= 8) {  
 val ratio = str1.length.toDouble / str2.length.toDouble  
 if (ratio > 2.0 || ratio < 0.5) 0.5  
 else 1.0  
 } else 1.0  
 EditDistance.sim(str1, str2).toDouble \* factor  
}  
  
/\*\*  
 \* 计算jaccard  
 \* @param segment1待计算的分词字符串1  
 \* @param segment2待计算的分词字符串2  
 \*/  
def calJaccardSimilar(segment1 : String, segment2 : String, separator : String) : Double = {  
 val seg1 = segment1.split(separator)  
 val seg2 = segment2.split(separator)  
 val inSeg = seg1 intersect seg2  
 val inLength = inSeg.length  
 val unionLength = seg1.length + seg2.length - inLength  
 inLength.toDouble / unionLength.toDouble  
}

### 获取文本关键词

读取文本数据，先用jieba进行分词，去除停用词。数据有id和content两列，id为文档id，content为文档内容。

val rdd = sc.textFile(dataPath).  
 map(\_.split(“\t”)).map(x=>(x(0),x(1)))

添加build.sbt中的配置项[[21]](#footnote-21)：

*libraryDependencies* += “com.huaban” % “jieba-analysis” % “1.0.2”

import com.huaban.analysis.jieba.JiebaSegmenter  
  
def jieba\_analysis(input: String): List[String] = {  
 val stopWords = “”.split(“\\|”, -1)  
 val segment = new JiebaSegmenter  
 val seg\_list = segment.sentenceProcess(input)  
 if (stopWords != null) {  
 for(sw <- stopWords){  
 if(seg\_list.contains(sw)) seg\_list.remove(sw)  
 }  
 }else *print*(“Stopwords is null”)  
 seg\_list.asInstanceOf[List[String]]  
}

使用java版的jieba分词。首先读取本地的停用词表stop\_words.txt，利用JiebaSegmenter对象进行分词，最后把分词的结果中的停用词过滤掉。因为jieba分词的结果类型为java.util.List，上面使用asInstanceOf强制转换成scala中的List类型。

以下是调用jieba\_analysis分词方法：

val splitWordRdd = rdd.map(file => {  
 (file.\_1, jieba\_analysis(file.\_2, “ “).split(“ “))  
})

Spark MLlib中提供了TF和IDF的方法，但是由于Spark提供的TF算法是不可逆的，也就是无法获取TF的结果对应的原始句子的文字，这里使用CountVectorizer代替[[22]](#footnote-22)。

CountVectorizer算法输出词频向量，使用IDF算法输出逆词频向量。

通过结果向量解析出关键词，cvModel.vocabulary中存储向量索引与原始句子的词的对应关系。

创建DataFrame的类型进行关键词提取：

val df = spark.createDataFrame(splitWordRdd).  
 toDF(“id”, “words”)  
  
val cvModel: CountVectorizerModel = new CountVectorizer().  
 setInputCol(“words”).  
 setOutputCol(“rawFeatures”).  
 setVocabSize(1000000). //向量长度  
 setMinDF(2). //词汇出现次数必须大于等于2  
 fit(df)  
val cvDf = cvModel.transform(df)

val idf = new IDF().  
 setInputCol(“rawFeatures”).  
 setOutputCol(“features”)  
  
 val idfModel = idf.fit(cvDf)  
 val idfData = idfModel.transform(cvDf)  
  
 val voc= cvModel.vocabulary  
 val keywordsDF = idfData.select(“id”,”features”).rdd.  
 map { x => {  
 val name = x.getAs[String](0)  
 //idf结果以稀疏矩阵保存  
 val v = x.getAs[org.apache.spark.ml.linalg.SparseVector](1)  
 var arrW = ArrayBuffer[String]()  
 var arrV = ArrayBuffer[Double]()  
 v.foreachActive((index:Int,value:Double)=>{  
 arrW += voc(index)  
 arrV += value  
 })  
 //根据idf值从大到小排序  
// (name, (arrW zip arrV).toList.sortBy(-\_.\_2).toArray)  
 (name, arrW.toArray)  
 }  
 }.toDF(“id”,”keywords”)

### 使用Spark LSH计算文本相似度

Spark获取文本的关键词，通过wordvec特征计算文本相似度[[23]](#footnote-23)。

其中Word2Vec即将词转换为词向量，这样词之间的关系就可以向量距离去定量计算，距离越近的两个词相似性也较高。Spark MLlib中文档的词向量，就是这个文档所有词的词向量的平均值[[24]](#footnote-24)。所以，我们使用提取的关键词来计算文档的词向量，因为原始文档的词向量直接时间的化会包含很多无意义的词。。

这里采用Word2Vec+LSH的方案，主要是因为Word2Vec将文档转换成了一个向量，接下来就是要求两篇文章的相似度，即其词向量的欧式距离，距离越近的则越相似。但是，海量数据的环境下，要对文档两两求求相似度再进行过滤。这个操作消耗的资源很大，并且时间相当长，通过spark-LSH来进行相似性检索极大的缓解这个压力。

下面是相关scala代码:

//Word2Vec获取关键词词向量  
val word2Vec = new Word2Vec().  
 setInputCol(“keywords”).  
 setOutputCol(“wordvec”).  
 setVectorSize(15).  
 setMinCount(0)  
val wvModel = word2Vec.fit(keywordsDF)  
val w2vDf = wvModel.transform(keywordsDF)  
w2vDf.show(false)  
  
val brp = new BucketedRandomProjectionLSH().  
 setBucketLength(4.0).  
 setNumHashTables(10).  
 setInputCol(“wordvec”).  
 setOutputCol(“hashes”)  
val brpModel = brp.fit(w2vDf)  
val tsDf = brpModel.transform(w2vDf)  
  
val brpDf = brpModel.approxSimilarityJoin(tsDf, tsDf, 0.015, “EuclideanDistance”)  
brpDf.show(false)  
  
import org.apache.spark.sql.\_  
import org.apache.spark.sql.functions.\_  
val getIdFun = udf((input:Row)=> {  
 input(0).toString  
})  
val corrDf = brpDf.withColumn(“id”,getIdFun(col(“datasetA”))).  
 withColumn(“id\_sim”,getIdFun(col(“datasetB”))).  
 drop(“datasetA”).drop(“datasetB”).drop(“EuclideanDistance”)  
corrDf.show(false)

## 基于图片的相似度

# 基于近邻的协同过滤

根据上一章的相似度计算可以发现，到目前为止并没有利用任何关于用户和物品的相关信息和特征。实际上，只要用户数量足够多，通过评分计算的相似物品推荐效果就已经相当好了。协同过滤（Collaborate Filer）算法正是利用了这一集体智慧的方法。

协同过滤算法（collaborate filter，下文简称CF）在推荐系统发展初期占据领导地位，即使在今天日益庞大的推荐系统架构中，CF算法也是必不可少的。CF算法主要分为两类，一类是基于领域的方法(neighborhood methods)也称为基于内存的（Memory\_Based Method），另一类是隐语义模型(latent factor models)，后者一个最成功的实现就是矩阵分解(matrix factorization)。

**Memory\_Based算法**

## 基于用户的协同过滤

基于用户的协同过滤，即User-Based CF (User-Based Collaborative Filtering)，是基于一个这样的假设“跟你爱好相同的人喜欢的物品，你很可能也喜欢”，所以User-Based CF主要的任务就是找出用户的最近邻居，从而根据最近邻居的喜好做出未知项的评分预测[[25]](#footnote-25)。  
User-Based CF算法可以分为4个步骤：数据表示、最近邻查询、评分预测、推荐结果产生。

根据上一章相似度计算介绍，我们知道对于一个评分矩阵：

|  |  |  |  |
| --- | --- | --- | --- |
| U\I | Item1 | Item2 | Item3 |
| User1 | 1 | 0 | 1 |
| User2 | 0 | 1 | 0 |
| User3 | 1 | 0 | 0 |

或者是：

|  |  |  |  |
| --- | --- | --- | --- |
| U\I | Item1 | Item2 | Item3 |
| User1 | 3 | 2 | 4 |
| User2 | 2 | 4 | 3 |
| User3 | 5 | 2 | 0 |

同理，在实际环境中，依次计算当前用户与其他所有用户的相似度，选取相似度最大的前K个用户作为该用户的最近邻集和。

### 显示反馈数据的评分预测

评分预测的方法也有多种，常用的是加权求和的方法计算预测值。将用户u的最近邻居用户对物品i的打分进行加权求和，权值为各个最近邻居用户与用户u的相似度，然后对加权求和的结果求平均，计算得到用户u对物品i的预测打分，公式如下：

N代表用户u的最近邻居用户集合，n是属于N中的一个用户， 表示用户u和用户n的相似度，表示用户n对物品i的评分。根据该公式逐个计算出用户u未评分的所有物品的预测评分，得到预测评分集合。

### 隐式反馈数据的过滤

不同于显式反馈数据，隐式反馈没有办法通过预测评分来获得推荐结果。

未完待续。

## 基于物品的协同过滤

基于物品（item-Based）的协同过滤跟上面基于用户的协同过滤算法原理基本相同，只是这里把相似度计算的对象转换为物品。

预测显式反馈数据的评分

根据物品相似度计算得到的（userID，itemId，sim），根据用户进行分组，取每组相似度最大的前K个物品的评分。最后按照预测值计算公式求出预测评分值[[26]](#footnote-26)。

隐式反馈数据获取推荐列表

基于内容的推荐（content-based recommendation）通过用户历史感兴趣的信息，抽象信息内容共性，根据内容共性推荐其他信息。

比如，如何通过基于内容的推荐，来对求职者A进行职位推荐？

答：简要步骤如下

找到用户A历史感兴趣的职位集合

找到职位集合的具化内容

抽象具化内容的共性内容

由这些共性内容查找其他职位，并实施推荐

典型步骤：收集用户历史数据，历史数据的feature化，根据历史数据的feature去找相似的item

关键点： 历史数据的feature化， item的feature化，相似度计算

典型应用：新闻网站，头条网站 （Common for recommending **text-based products** (webpages, usenet news messages)

也可以是model-based，The user model can also be a **classifier** based on whatever technique (Neural Networks, Naïve Bayes...)

### 基于用户profile的推荐

通过用户的profile， 结合item的feature (包含item的profile)，进行推荐

用户的profile：人口基本特征/兴趣/偏好/用户等级，忠诚度/keyword， 可以是用户填写的， 也可以是计算/推测出来的

特点：不需要用户的历史数据，可以作为冷启动的解决方案之一。例如某些阅读应用，在用户第一次登录的时候，会要求用户选择感兴趣的domain

典型步骤：各种途径获得user profile，计算user profile与item feature之间的相似度，推荐

与“基于内容的推荐”的区分和联系

基于用户的内容，可以生产出profile的内容（比如关键词，tag，分类等等）

有时并不严格区分 （content-based user profile）

问题：可信度不高，因为即便是年龄、性别等属性都相同的用户，也很有可能对物品有截然不同的偏好；可解释性不高

还有一种广泛使用的更有实际意义的user profile：考虑该User打过分的所有Item，将这些Item的Item Profile的每一项分别进行加权平均，得到一个综合的Profile，作为该用户的User Profile。这种User Profile的优点是非常容易计算其与Item之间的相似度，同时比较准确地描述了该用户在Item上的偏好，巧妙地避开了用户私人信息这一很难获得的数据，具有保护隐私的能力，进一步，如果加入时间因素，还可以研究用户在Item上偏好的变化等等，因此受到广泛应用

客户特征的向量空间模型

图片包含 文字

描述已自动生成

将上文抽取的用户属性和标签进行离散化，重要特征设置相应的权重，然后将每个客户的特征向量化。通过计算新客户和已成交客户之间的相似度和筛选条件，在所有新客户中挖掘出高度潜在客户。

### 用户行为权重模型

很多开卡和未开卡的客户在 “掌上生活”APP中有大量的点击、浏览、查询等行为，而开卡用户的刷卡消费记录、分期情况，都可以反映客户对于保险产品的潜在意向。通过这些行为的建模匹配其对于保险产品的兴趣度，从而进行有针对性的营销。

图片包含 墙壁

描述已自动生成

比如，根据用户的具体行为与保险的相关度进行评分，然后构建所有用户关于行为的评分矩阵，计算每个客户对于保险的喜好度。如果加入保单的维度，就可以算出每个客户对于每个保险产品的喜好度。

### 协同过滤

图片包含 屏幕截图

描述已自动生成

基于用户的协同过滤（user-based CF，给本客户推荐与其情况相似的其他客户普遍持有的保险产品）：比如上图用户1和用户3都购买了悦享康健-重大疾病险和泰康家倍-家庭医疗险，可以根据用户评分计算出两者相似度较高，所以把用户3买过的诺享无忧-意外险推荐给用户1。

基于用户的协同过滤也可以通过neighborhood\_based算法实现：设定相应的阈值找出跟用户1距离相近的其他用户，他们购买过的保险具有群体相似性。

基于物品的协同过滤（item-based CF）：计算物品之间的行为或特征相似度，再进行推荐。比如，给购买过住院医疗险的客户推荐重大疾病险，因为这两种保险同现频率很高。此处的物品指的就是保险产品。

# 基于模型的协同过滤

## 隐因子模型（Latent Fator Model）

协同过滤的思路就是基于用户和物品的交互行为，要么计算用户间的相似度，推荐相似度高的用户喜欢的物品，因为这两个用户可能兴趣相投；要么就是计算物品间的相似度，推荐和历史记录相似度很高的物品，因为他们可能属于同一类别的商品。我们做决策的基础都是默认了商品是有类别的，可能有的用户都喜欢某一类商品，所以这些用户之间相似度高，可能有的商品是属于同一类别的，因此这些商品的相似度很高。那既然这样，有没有可能直接得到商品的类别呢？这样我们就可以直接根据类别去进行推荐了~

假设一个用户-物品的矩阵：

实际情况下，用户不可能什么商品都买，所以，该矩阵必然是一个稀疏矩阵，任意一个矩阵必然可以被分解成2个矩阵的乘积：



=

根据上述公式，k就是潜在因子的个数，举个例子，你去买衣服，你可能买了裙子，露背装，我去买衣服，我买了牛仔裤，潮牌T恤，影响我们购买商品的差异的原因可能有很多点，但是必然有些原因占比重要些，比如性别，收入，有一些可能不那么重要比如天气，心情。而拆分成的Pu矩阵表示了这些潜在因子对我或者你的影响程度，Qi矩阵表示了各种商品对这些潜在因子的影响程度。

当我们尽可能的通过拆分矩阵的形式，目标使得拆分后的两个矩阵的乘积最匹配最上方的用户商品矩阵的已知的数据值，从而可以通过这两个矩阵的乘积填补掉空缺的值。

一个简单的评分预测方法：

使用向量bi表示电影i的评分相对于平均评分的偏差，向量bu表示用户u做出的评分相对于平均评分的偏差，将平均评分记做μ。

新的得分计算方式如下：

准备引入了商品及用户的实际分布的情况，有效的降低在测试数据上面的效果[[27]](#footnote-27)。

LFM就是基于这样的想法，假设商品存在若干个种类，那么每个用户对每个类会有一个兴趣度，同样的，每个类内又有若干种商品，每个商品在这个类内又会有一个对应的权重。这样，对于任何一个用户-商品对，我们都可以用下述公式来表达

图片包含 屏幕截图

描述已自动生成

假设一些用户，或者几个产品属于一个类型或者说是因子（factor），同时假设用户对于特定的因子有一定的喜好度。可以抽象出一个隐形因子空间，然后把用户和物品分别投影到这个空间上，来直接寻找用户-物品的喜好度。这个模型被称为隐因子模型。

下一步就是去计算矩阵P和Q，最常见的方法就是在训练集上不停迭代更新参数直至参数收敛。既然需要迭代，那我们得有一个损失函数或者目标函数作为我们迭代的依据，这里我们选用的是最常见的预测值和实际值差值的平方和，并且加上了正则项防止过拟合，具体如下[[28]](#footnote-28)：

熟悉机器学习的同学应该很清楚，接下来就是基于梯度下降去优化这个损失函数，这里就不赘述了。但是我们还有一个问题，对于显式反馈的数据，我们直接用评分数据作为训练集，但是对于隐式反馈而言，只有正样本，我们可以把这些数据标注为1，但同时我们也需要负样本，所以往往需要我们从所有样本里选择样本构建负样本集。这里在构建负样本的时候有一个小小的trick，要尽量选择那些非常热门但是用户却没有发生交互的物品，因为这样构建的样本往往更具有代表性，也有利于我们的training。在movielens数据集中，我们以样本出现的次数作为权重，随机选择样本构建负样本集，实现如下

def Random\_Negative\_Sampling(self, items):

ret = {}

for i in items:

ret[i] = 1

count = 0

for i in range(0, len(items)\*5):

item = self.items\_pool[random.randint(0, len(self.items\_pool)-1)]

if item in items:

continue

ret[item] = 0

count += 1

if count > 2\*len(items):

break

return ret

如何去计算一个用户对一个物品的喜爱程度？是不是直接用上面那个公式，用用户对类的喜爱程度和该物品在此类中的权重乘积的累加呀。那么恭喜你，如果这样做，效果非常不好，一是太难收敛，二是预测效果总是不好。为啥呢？实际预测的矩阵里只有0和1两个值，而Preference的计算则是一个实数，在梯度下降的时候要用到这两者的差值，而这个值变动的范围太大，一方面导致你基本很难收敛，另一方面超参数的选择是个非常头疼的问题。所以，在计算喜好度的时候，一定要给结果加Sigmod函数。

这样把你的结果约束到0和1之间，防止在计算误差的时候出现太大的波动从而影响收敛。这里我在movielen的数据集上也进行了实验，《推荐系统实战》那本书里选择的隐类个数为100，本来是想复现他的结果，不过我发现算起来实在是太慢了，所以我这里选择隐类F=10，学习率alpha=0.03，正则项稀疏lambda=0.01，在所有的用户上迭代了30轮次，最后在测试集上的表现如下

召回率6.91% 精确率23.2% 覆盖率40.8% 流行度5.20。

按照常理和书上所说，基于学习的这种方法的表现应该比基于统计的协同过滤表现要好点，但是这里的结果明显比之前的最好结果要差些，主要原因可能是这里的隐类选的太少了，书中的实验实际选择的隐类数目为100。但我发现隐类100的时候不仅每一轮计算的速度大大提升，并且收敛的速度也远远慢于隐类数目较少的时候，这也很容易理解，毕竟对应的需要优化的参数数目多了10倍，因此需要的训练样本也要更多，收敛自然更慢，有兴趣的或者有时间的同学可以去尝试一下。

然后在整个实现LFM的过程中，还有其他几个小地方大家可以注意一下

参数的初始化。所有参数初始化为0好像不是一个好选择，总是无法收敛。一般的做法都是选择在-1到1之间的随机数。

构建负样本和正样本的比例。书中说这是一个影响模型效果很重要的参数，实际上可以这么理解，相同轮次的迭代，负样本比例越高，其实越多的训练数据，自然模型的表达能力可能会更好。

超参数的选择。在实际使用的过程中，我发现超参数的选择对于LFM模型的表现真实至关重要，学习率过高直接导致没法收敛，但是太低了收敛的速度又太慢了，所以超参数的选择真的是一个非常头疼的事。

LFM的计算速度是真的很慢。每一轮迭代，要在每个用户上迭代其所有的行为记录和我们构造的负样本，导致计算真的非常耗时。而隐类个数的增加不仅导致需要优化的参数增加，对于数据的需求同样提高，这样直接导致收敛速度变得非常非常慢。

最后是对LFM算法的一些思考，它的本质其实就是矩阵分解，将之前的用户物品矩阵分解成了用户-隐类矩阵和隐类-物品矩阵，然后目标是让这两个矩阵的乘积与原矩阵的残差尽可能得小，同时引入了一些正则项防止过拟合。后来我发现其实LFM算法就是FunkSVD算法，是典型的基于矩阵分解的方法，后来又出现了基于它进行改进的方法。

## SVD解决矩阵分解问题

### SVD用于数据降维

首先，线代或者高等代数里面告诉我们：一个向量可以通过左乘一个矩阵的方式来进行拉伸，旋转，或者同时拉伸旋转。

所以，无论什么矩阵M，我们都可以找到一组正交基v1、v2，使得Mv1、Mv2也是正交的，不妨记其方向为为μ1、μ2。

Mv1=δ1u1；

Mv2=δ2u2；

存在向量x，在v1、v2空间里的表示为：x=（x·v1）v1+（x·v2）v2，

所以有，Mx有：

Mx=（x·v1）Mv1+（x·v2）Mv2

Mx=（x·v1）δ1u1+（x·v2）δ2u2

Mx=δ1u1(v1.T)x+δ2u2(v2.T)x

M=δ1u1(v1.T)+δ2u2(v2.T)

所以就有了那个非常有名的公式：

M=UΣV.T

U是有一组正交基构成的，V也是有一组正交基构成的，Σ是由δ1、δ2构成的，几何意义上来说，M的作用就是把一个向量由V的正交空间变换到U的正交空间上，而通过Σ的大小来控制缩放的力度。

我们还需要知道一些简单的推论，

通过MM.T，我们知道，δ的平方是MM.T的特征值

奇异值δ的数量决定了M=UΣV.T的复杂度，而奇异值的大小变化差异程度很大，通常前几个奇异值的平方就能占到全部奇异值的平方的90%，所以，我们可以通过控制奇异值的数量来优化原始矩阵乘积，去除掉一下噪声数据[[29]](#footnote-29)。

当你有一个多维度稀疏矩阵，通过矩阵因式分解你能够将用户-项目矩阵（user-item matrix）重构成低评分结构（low-rank structure），并且你能够通过两个低评分（ low-rank）矩阵相乘得出这个矩阵，其中矩阵的行包含潜在向量。

通过低评价矩阵乘积尽可能调整这个矩阵近似原始矩阵，以填充原始矩阵中缺失的项。

优点：更好解决可扩展性和稀疏问题而被广泛用于推荐系统

缺点：矩阵分解时间复杂度高，可采用梯度下降的方法减少计算复杂度

利用SVD解决矩阵分解（MF）问题任意一个M\*N的矩阵A（M行\*N列，M>N），可以被写成三个矩阵的乘积：

U：（M行M列的列正交矩阵）

S：（M\*N的对角线矩阵，矩阵元素非负）

V：（N\*N的正交矩阵的倒置）

即A=U\*S\*V’（注意矩阵V需要倒置）

简单总结就是选取S对角阵中的前k个元素即可对U,S进行降维，利用，令U=U\*S, 则U\*V’可以近似还原并填充原矩阵?【这句话我认为不对的吧。还原是近似接近原矩阵， 如果原来是0，即未评分，还原的后的矩阵应该还是很接近0才对】，应该采用后面的方法对未评分的元素进行预测。

### SVD进行评分预测

根据上一节的内容，我们知道评分矩阵R可以用两个 矩阵P和Q的乘积来表示：

其中，U表示用户数，I表示商品数，K就是潜在因子个数。

|  |  |  |  |
| --- | --- | --- | --- |
| 评分矩阵（m\*n） | | | |
| 用户\电影 | 钢铁侠3 | 我不是药神 | 复仇者联盟3 |
| user1 | 4 | 3 | 5 |
| user2 | 2 | 5 | 4 |
| user3 | 2 | 4 | 0 |

将表格中的内容转化成一个m\*n的矩阵R:

首先可以通过那些已知数据的来训练这两个乘积矩阵。那么位置的评分也就可以用P的某一行乘以Q的某一列得到：



这时预测用户u对商品i的评分，它等于P矩阵的第u行乘以Q矩阵的第i列。这时最基本的SVD算法，下面我们来看如何确定和。

假设已知的评分为，则真实值与预测值之间的误差为：

由此可以计算均方误差：

只要通过训练把MSE降到最小那么P、Q就能最好地拟合R了。常规的来讲，梯度下降是非常好的求解方式，常见的包括随机梯度下降，批量梯度下降。

随机梯度下降一定程度会避免局部最小但是计算量大，批量梯度计算量小但是会存在鞍点计算误区的问题。

先求得MSE在变量（也就是P矩阵的第u行第k列的值）处的梯度：

现在得到了目标函数在处的梯度了，那么按照梯度下降法，将往负梯度方向变化。令更新的步长（也就是学习速率）为。

那么的更新公式为：

同理，得到的更新公式为：

**RSVD**

根据上面的公式，很显然这样去求矩阵QP必然会存在过拟合的问题，导致对实际数据预测时效果远差于训练数据，我们可以增加正则项。

对所有的变量加入正则惩罚项，重新计算梯度如下：

这就是正则svd，也叫做Rsvd，也是我们用的比较多的svd的方法。

**偏移RSVD**

在最开始讲了，Koren在NetFlix大赛里面除了考虑了对原始数据的拟合情况，也考虑了用户的评分、商品的平均得分相对于整体数据的偏移情况，有了新的得分公式：Rui＝μ+bi+bu，影响的只有eui，后面的Pu、qi的正则不受影响，但是新增了bi、bu的正则项，重新计算每一项的偏导数：

bu、bi的更新式子[[30]](#footnote-30)：

其余的都不发生改变，这就叫做有偏移下的Rsvd

无论是Rsvd还是有偏移的Rsvd，当原始的用户对商品的评分矩阵过大，比如有3亿用户，3亿商品，形成9亿商品集合的时候，这就是一个比较不可能完成的存储任务，而且里面绝大多数都是0的稀疏矩阵。

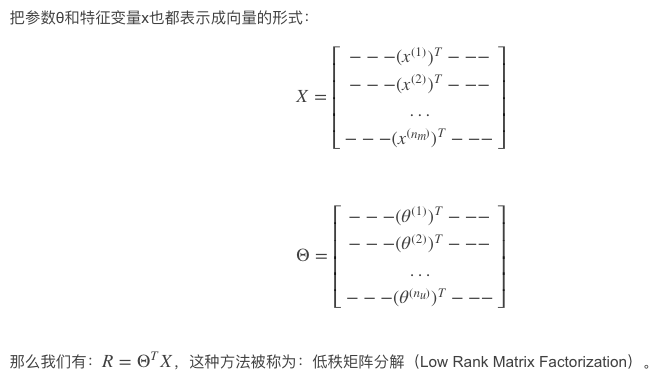
**Asvd及Svd++**

这边，我们引入两个集合：R(u)表示用户u评过分的商品集合，N(u)表示用户u浏览过但没有评过分的商品集合，Xj和Yj是商品的属性。

Asvd的rui的评分方式：

Svd++的rui的评分方式：

无论是Asvd还是Svd++，都干掉了原来庞大的P矩阵，取而代之的是两个用户浏览评分矩阵大大缩小了存储的空间，但是随着而来的是一大把更多的未知参数及迭代的复杂程度，所有在训练时间上而言，会大大的增加。



由此找电影i相似的电影j：可以计算两个特征向量的距离，其中距离最小的就是最相似的电影

把右上表中的第二列（col=《我不是药神》）和第三列（col=《复仇者联盟3》）分别与

右下表的第二列（col=User2）对应元素相乘后相加（即点乘），得到的结果分别是（55 + -40=25）和（-25 + 50=-10），

这两个分数的物理意义分别是李四喜欢《我不是药神》和李四喜欢《变形金钢4》的程度，

可以看出李四更喜欢《我不是药神》。那么我们把右上表的每列和右下表的每列各元素相乘后相加，

是不是可以得到每个人喜欢每部片子的程度呢？这不就是我们左边Rating matirx图表示的意思嘛[[31]](#footnote-31)！

我们把评分矩阵（Rating Matrix）计作V， 𝑉∈𝑅𝑛×𝑚，那么V的每一行𝑉𝑖代表一个人的所有评分，每一列𝑉𝑗代表某一部电影所有人的评分，𝑉𝑖𝑗代表某个人i对某部电影j的评分。对应电影推荐来说，V必定是稀疏的，因为电影数量（列的数目）是巨大的，V中必定有很多很多项为null。

我们接下来看这两个矩阵U（Users Features Matrix ）和M（Movie Features Matrix）。U为用户对特征的偏好程度矩阵，M为电影对特征的拥有程度矩阵。𝑈∈𝑅𝑓×𝑛的每一行表示用户，每一列表示一个特征，它们的值表示用户与某一特征的相关性，值越大，表明特征越明显。矩阵𝑀∈𝑅𝑓×𝑚，的每一行表示电影，每一列表示电影与特征的关联。

那么U和M怎么得到呢？

还记得SVD的公式嘛？𝑉=𝑈𝑀𝑇，其实公式右边的中间还有个对角矩阵S，我们可以把他看成跟U相乘合并（The S matrix is left blended into the feature vectors, so only U and M remain）。其实，U和M这两个矩阵是通过学习的方式得到的，而不是直接做矩阵分解。我们定义如下的损失函数：

这里的评价指标metric采用的是RMSE。其中𝑝(𝑈𝑖,𝑀𝑗)代表用户i对电影j的预测，最常用的预测函数p就是点乘，即 𝑝(𝑈𝑖,𝑀𝑗)=𝑈𝑇𝑖𝑀𝑗。式(1)中的 𝐼∈{0,1}𝑛×𝑚，为一个指示器，指示相应位置是否有评分，有评分为1，没有为0。等式右边最后两项是正则化项，防止过拟合，这里不过多展开。

到这里，已经变成了一个机器学习的常见问题了，即最小化损失函数，用得最多的优化方法就是梯度下降，当然还有很多梯度下降的变体。下图是简单的梯度下降算法。

添加偏置项

梯度下降方法

: 训练集中所有记录的评分的全局平均数。在不同网站中，因为网站定位和销售的物品不同，网站的整体评分分布也会显示出一些差异。比如有些网站中的用户就是喜欢打高分，而另一些网站的用户就是喜欢打低分。而全局平均数可以表示网站本身对用户评分的影响。

用户偏置（user bias）项。这一项表示了用户的评分习惯中和物品没有关系的那种因素。比如有些用户就是比较苛刻，对什么东西要求都很高，那么他的评分就会偏低，而有些用户比较宽容，对什么东西都觉得不错，那么他的评分就会偏高。

: 物品偏置（item bias）项。这一项表示了物品接受的评分中和用户没有什么关系的因素。比如有些物品本身质量就很高，因此获得的评分相对都比较高，而有些物品本身质量很差，因此获得的评分相对都会比较低。

### 常见的LFM算法

SVD奇异值分解

这里的SVD推荐本质上是model-based，跟传统数学意义的SVD没有太大关系，只不过借鉴了SVD分解R=U∗S∗VR=U∗S∗V这个形式，通过最优化方法进行模型拟合，求得R=U∗VR=U∗V。

BiasSVD。在FunkSVD的基础上又引入了平均得分、用户偏置项和物品偏置项，相比于FunkSVD又多了两个需要优化的参数，但优化方法啥的都是一样的。事实表明，由于考虑了这些偏置项，令BiasSVD在某些场景下变现会优异很多。

SVD++。在BiasSVD的基础上又引入了用户的隐式反馈，也就是用户之前如果对该物品产生过交互，则对它的评分要进行修正。结构越来越复杂，计算也会越来越慢，虽然说SVD++在竞赛中大放异彩，但实际应用比较困难，计算代价太大，无法做到实时推荐，而且可解释性也不强。

### LFM与基于邻域的方法对比：

图片包含 屏幕截图

描述已自动生成

## 基于矩阵分解的LFM算法应用

### LibMF

矩阵分解通常用于显示反馈数据的推荐系统。

LIBMF使用潜在空间中两个矩阵的乘积，近似得到一个不完全矩阵。它可以解决实值矩阵分解、二元矩阵分解和一类矩阵分解问题。我们使用LibMF进行矩阵分解和评分预测，主要考虑到它的运算速度相当快。一方面支持在多核机器中并行计算，并通过CPU指令加速向量运算；另一方面，如果数据量过大，可以进行磁盘训练，减少内存使用。

**安装：**

下载libmf-2.01的安装文件[[32]](#footnote-32)。解压后进入目录：

执行make命令，发现目录下多了mf-train和mf-predict这两个可执行文件。

图片包含 物体

描述已自动生成

**数据格式**

<row\_idx> <col\_idx> <value>

在demo目录中，文件real\_matrix.tr.txt’ 和 ‘real\_matrix.te.txt’是真值矩阵分解real-valued matrix factorization (RVMF)演示的训练和[测试](http://lib.csdn.net/base/softwaretest" \o "软件测试知识库)数据集。二元矩阵分解binary matrix factorization (BMF)中，‘binary\_matrix.tr.txt’ 和‘binary\_matrix.te.txt.’中<value>集是{-1, 1}。在一类矩阵分解（one-class MF）中，所有的<value>都是正的。

**模型格式**

LibMF把一个训练矩阵R变为一个k-by-m的矩阵‘P’和一个k-by-n的矩阵‘Q’，也就是R近似于P’Q。训练过程·结束后，这两个因子矩阵P和Ｑ被存到一个模型文件中。这个文件以如下打头：

‘f’: f表示MF矩阵分解问题的损失函数，

‘m’: 矩阵R行的数目,  
   ‘n’: 矩阵R列的数目,  
   ‘k’: 隐因子的数目,  
     ‘b’: 矩阵R所有元素的平均值。

从第五行开始，Ｐ和Ｑ的列就被一行接一行的存储。每一行，都有两个领导标志跟在一列值后面。第一个标志是被存储列的名字，第二个标志表明值的类型。如果第二个标志是‘T’，列是真值。否则，列的所有值是NaN。举个例子：

并且b=0.5，则模型文件的内容是：

    --------model file--------  
       m 3  
       n 2  
       k 3  
       b 0.5  
       p0 T 1 3 5  
       p1 F 0 0 0  
       p2 T 2 4 6  
     q0 T -1 -3 -5  
       q1 T -2 -4 -6  
       --------------------------

**命令使用**

**mf-train**

用法: mf-train [options] training\_set\_file [model\_file]

参数选择：

* l1 <lambda>,<lambda>: set L1-regularization parameters for P and Q.

(default 0) If only one value is specified, P and Q share the same

lambda.

* l2 <lambda>,<lambda>: set L2-regularization parameters for P and Q.

(default 0.1) If only one value is specified, P and Q share the same

lambda.

* f <loss>: 指定损失函数 (default 0)

for real-valued matrix factorization

0 -- squared error (L2-norm)

1 -- absolute error (L1-norm)

2 -- generalized KL-divergence (--nmf is required)

for binary matrix factorization

5 -- logarithmic error

6 -- squared hinge loss

7 -- hinge loss

for one-class matrix factorization

10 -- row-oriented pair-wise logarithmic loss

11 -- column-oriented pair-wise logarithmic loss

* k <dimensions>: 设置维度的数量，也就是隐因子数量 (default 8)
* t <iter>: 设置迭代次数 (default 20)
* r <eta>: 设置初始学习率(default 0.1)
* s <threads>: set number of threads (default 12)
* n <bins>: set number of bins (may be adjusted by LIBMF for speed)
* p <path>: set path to the validation set
* v <fold>: set number of folds for cross validation
* quiet: quiet mode (no outputs)
* nmf: perform non-negative matrix factorization
* disk: perform disk-level training (will create a buffer file)

“mf-train”是LibMF最主要的训练命令。每次迭代，下列信息都被打印出来：

* iter: 迭代的索引   
    - tr\_xxxx: xxxx is the evaluation criterion on the training set  
      - va\_xxxx: the same criterion on the validation set if ‘-p’ is set  
      - obj: objective function value

这里的‘tr\_xxxx’ 和 ‘obj’ 都是估计的，因为计算真的值太耗时间了。  
对于不同的损失，标准如下：

        <loss>: <evaluation criterion>  
        -       0: root mean square error (RMSE)  
        -       1: mean absolute error (MAE)  
        -       2: generalized KL-divergence (KL)  
        -       5: logarithmic loss  
        -   6 & 7: accuracy  
        - 10 & 11: pair-wise logarithmic loss (BprLoss)

**mf-predict**

用法：mf-predict [options] test\_file model\_file output\_file

options:

* e <criterion>: set the evaluation criterion (default 0)

0: root mean square error

1: mean absolute error

2: generalized KL-divergence

5: logarithmic loss

6: accuracy

10: row-oriented mean percentile rank (row-oriented MPR)

11: colum-oriented mean percentile rank (column-oriented MPR)

12: row-oriented area under ROC curve (row-oriented AUC)

13: column-oriented area under ROC curve (column-oriented AUC)

**案例**

默认参数训练一个模型：

* mf-train real\_matrix.tr.txt model

训练具有正则化系数的模型，系数如下：

* mf-train -l1 0.05 -l2 0.01 real\_matrix.tr.txt model

coefficient of L1-norm regularization on P = 0.05

coefficient of L1-norm regularization on Q = 0.05

coefficient of L2-norm regularization on P = 0.01

coefficient of L2-norm regularization on Q = 0.01

另一组正则化系数，训练模型：

* mf-train -l1 0.015,0 -l2 0.01,0.005 real\_matrix.tr.txt model

coefficient of L1-norm regularization on P = 0.05

coefficient of L1-norm regularization on Q = 0

coefficient of L2-norm regularization on P = 0.01

coefficient of L2-norm regularization on Q = 0.03

* mf-train -f 5 -l1 0,0.02 -k 100 -t 30 -r 0.02 -s 4 binary\_matrix.tr.txt model

直接预测，输出评分到指定文件：

* mf-predict real\_matrix.te.txt model output

预测并输出MAE值：

* mf-predict -e 1 real\_matrix.te.txt model output

### LibFM

这一类模型的共同特点是矩阵分解。即对用户-物品评分矩阵分解成若干个小矩阵，目的是分解之后的矩阵乘积接近原始矩阵，于是也实现了对原始矩阵为空的值的预测。

在这些方法中，比较重要的几个参数有：隐特征个数，随机梯度下降中的学习率，正则化参数，总迭代次数等。具体在每个方法中这些参数的最优值也不尽相同。

### ALS最小二乘法

最小二乘法（ALS，alternating least squares）是一种求解矩阵分解问题的优化方法。它功能强大而且容易实现并行化。所以它比较适合像Spark这样的的大数据平台。

ALS的实现原理是迭代式求解一系列最小二乘法回归问题。在每一次迭代时，固定用户因子矩阵或物品因子矩阵中的一个，然后用固定的这个矩阵以及评分数据来更新另一个矩阵。之后，被更新的矩阵被固定住，再更新另一个矩阵。如此迭代，直到模型收敛。

由于q和p未知，目标函数是非凸函数。但如果假设其中一个当前值固定，则该优化问题可解。ALS便通过交替上述假设来迭代求解。

* rank:对应ALS模型中的因子个数，也就是在低阶近似矩阵中的隐含特征个数。因子个 数一般越多越好。但它也会直接影响模型训练和保存时所需的内存开销，尤其是在用户 和物品很多的时候。因此实践中该参数常作为训练效果与系统开销之间的调节参数。通 常，其合理取值范围为10~200。
* iterations:对应运行时的迭代次数。ALS能确保每次迭代都能降低评级矩阵的重建误 差，但一般经少数次迭代后ALS模型便已能收敛为一个比较合理的好模型。这样，大部 分情况下都没必要迭代太多次(10次左右一般就挺好)。
* numBlocks:对应用户和物品将分为将并行计算的块数(默认为10)。该数取决于集群的 节点数和输入分块的方式。
* regParam:对应ALS的正则化参数(默认为1.0)。常数λ被称为正则化参数。本质上， 当用户或物品矩阵的规模过大时，它会减小矩阵的因子。这对数值稳定很重要，而且总 会引入某种正则化方法。
* implicitPrefs:对应标识矩阵中的值是隐式反馈值还是显式反馈值，两种分别会用ALS隐式反馈衍生模型(ALS-WR)和显示反馈衍生模型来建模。默认为False，即显式 反馈值。
* alpha:这是ALS隐式反馈模型的一个参数，它确定了反馈对应的基准可信度，默认为1.0。
* nongeative:指定是否使用非负约束条件。默认为false。

读取MovieLens评分数据ratings.csv：  
val moviePath = “/data/ml-1m/ratings.csv”  
  
val ratingsDF = spark.read.  
 option(“header”, “true”).  
 option(“delimiter”, “,”).  
 format(“csv”).  
 load(moviePath)  
  
显示一条数据  
ratingsDF.first()

输出如下：

org.apache.spark.sql.Row = [1,1193,5,978300760]

生成训练数据，将ratingDF由DataFrame转换为Rating(user,product,rating)格式的Dataset类型：

import org.apache.spark.sql.types.{DoubleType, IntegerType}

import org.apache.spark.mllib.recommendation.Rating  
val trainData = ratingsDF.select(“userId”, “movieId”, “rating”).  
 withColumn(“user”, ‘userId.cast(IntegerType)).  
 withColumn(“product”, ‘movieId.cast(IntegerType)).  
 withColumn(“rating”, ‘rating.cast(DoubleType)).as[Rating]  
  
训练模型  
进行ALS推荐系统模型训练。MLlib中的ALS算法接收三个参数：  
- rank：对应的是隐因子的个数，这个值设置越高越准，但是也会产生更多的计算量。一般将这个值设置为10-200；  
- iterations：对应迭代次数，一般设置个10就够了；  
- lambda：该参数控制正则化过程，其值越高，正则化程度就越深。一般设置为0.01。  
  
首先，执行以下代码，启动ALS训练：  
  
import org.apache.spark.ml.recommendation.ALS  
val model = new ALS().  
 setImplicitPrefs(true).  
 setRank(10).  
 setRegParam(0.01).  
 setAlpha(1.0).  
 setMaxIter(5).  
 setUserCol(“user”).  
 setItemCol(“product”).  
 setRatingCol(“rating”).  
 setPredictionCol(“prediction”).  
 fit(trainData)  
  
返回类型为ALSModel对象，它将结果分别保存到两个(id,factor)RDD里面，分别名为userFactors和productFactors。选取一个显示:  
model.userFactors.first()

输出为：

org.apache.spark.sql.Row = [10,WrappedArray(-0.67133933, -1.0876986, -3.481903, -1.3368399, -1.0553768, -0.78186196, 0.3581914, -0.39093563, 0.3642208, 0.51764965)]  
  
使用ALS模型进行预测  
预测用户1对物品123的评分  
val predictRating = model  
  
为用户1推荐10个物品  
val userId = 1  
val N = 10  
  
import org.apache.spark.ml.recommendation.ALSModel  
def makeRecommendations(model: ALSModel, userID: Int, howMany: Int): DataFrame = {  
 val toRecommend = model.itemFactors.  
 select($”id”.as(“product”)).  
 withColumn(“user”, *lit*(userID))  
 model.transform(toRecommend).  
 select(“product”, “prediction”).  
 orderBy($”prediction”.desc).  
 limit(howMany)  
}

如果报错org.apache.spark.sql.AnalysisException: Detected implicit cartesian product for LEFT OUTER join between logical plans  
Spark 2.x中默认不支持笛卡尔积操作，需要通过参数spark.sql.crossJoin.enabled开启  
spark.*conf*.set(“spark.sql.crossJoin.enabled”, “true”)  
  
val topNForUser1 = makeRecommendations(model, userId, N)  
  
查看给User1推荐的结果  
topNForUser1.show()  
  
给所有用户推荐TopN个电影  
model.recommendForAllUsers(N)  
  
  
**检验推荐效果**  
导入电影数据movies.csv:  
def buildItemAlias(itemAlias: RDD[String]): Map[Int, Int] =  
 itemAlias.filter(!\_.contains(“movie”)).  
 flatMap { line =>  
 val tokens = line.split(‘\t’)  
 if (tokens(0).isEmpty) {  
 None  
 } else {  
 Some((tokens(0).toInt, tokens(1).toInt))  
 }  
 }.collectAsMap()  
  
val titles = buildItemAlias(spark.sparkContext.textFile(“ml-1m/movies.csv”))

获取某用户的所有观影记录并打印：

建立用户名-其他RDD，并仅获取用户789的记录

val moviesForUser = trainData.filter(\_.user == 1)

获取用户评分最高的10部电影，并打印电影名和评分值

moviesForUser.sort(“rating”).take(10).

map(rating => (titles(rating.product), rating.rating)).foreach(println)

物品推荐

实际场景中，需要找到一个物品的相似物品

导入jblas库中的矩阵类

import org.jblas.DoubleMatrix

定义余弦相似度方法

def cosineSimilarity(vec1: DoubleMatrix, vec2: DoubleMatrix): Double = {

vec1.dot(vec2) / (vec1.norm2() \* vec2.norm2())

}

设置推荐个数为K，选择itemId为1的作为实验。∞

val K = 10

val itemId = 1

获取该物品的隐因子向量，并它与所有物品的隐因子向量的相似度。

val featList = model.itemFactors.

filter($”id”=== itemId).

select(“features”).as[Array[Double]].collect().head

val itemFactor = new DoubleMatrix(featList)

case class VectorFeat(id: Int, features: Array[Double])

val sims = model.itemFactors.as[VectorFeat].

map{ row =>

val factorVector = new DoubleMatrix(row.features)

val sim = cosineSimilarity(factorVector, itemFactor)

(row.id, sim)

}

得到RDD([String, Double])的计算结果，按照相似度大小进行排序，打印结果。

val sortedSims = sims.rdd.

top(K)(Ordering.by[(Int, Double), Double] { case (id, similarity) => similarity })

println(sortedSims.mkString(“\n”))

输出：

(1,1.0)

(34,0.9170793265884063)

(588,0.8899633317913926)

(3114,0.8211203043339788)

(1265,0.8210150099463293)

(2081,0.8019215081943607)

(2791,0.7984530251436968)

(1073,0.790919145627971)

(1259,0.7848895544665173)

(1197,0.7660933410852737)

**参数设置**

通过计算AUC进行参数寻优，利用GridSearch查找rank、regParam、alpha的最优参数。

在最开始的时候我们随机抽一批数据作为验证集。

事实上通常数据被分为三个子集：训练集、验证集（Cross-Validation）和测试集。

**效果评价**

在Spark的ALS推荐系统中，最常用到的两个推荐指标分别为MSE和MAPK。其中MSE就是均方误差，是基于评分矩阵的推荐系统的必用指标。那么MAPK又是什么呢？

它称为K值平均准确率，最多用于TopN推荐中，它表示数据集范围内K个推荐物品与实际用户购买物品的吻合度。具体公式请读者自行参考有关文档。

推荐系统就是一个[基于用户-物品评分矩阵的TopN推荐系统]，下面步骤分别用来获取本文推荐系统中的这两个指标。

1. 计算MSE和RMSE

val predictions = model.transform(testData)  
println(predictions.printSchema())  
  
import org.apache.spark.ml.evaluation.RegressionEvaluator  
val evaluator = new RegressionEvaluator().  
 setMetricName("rmse").  
 setLabelCol("rating").  
 setPredictionCol("prediction")  
val rmse = evaluator.evaluate(predictions)  
  
println(s"Root-mean-square error = $rmse")

2. 计算MAPK

### Spark-GraphX SVD++

第二种主流方法是去挖掘一些隐性变量（ latent variables ），避免了第一种方法需要找到与目标用户准确匹配的其他用户的要求。这听起来会有点晦涩难懂，但它的基本原理并不复杂，如图 7.4 所示 。 通过隐性变量 ， 我们就可 以使用 一个向量来表示每一部影片 ，向量代表电影拥有的不同特性。我们 的例子使用了两个隐性变量，这样每部电影就可以使用一个二维向量来表示。尽管从图 7.4 看来 ， 《星球大战》只具有科幻电影这一个特性，但是我们依然用 一个长度为 2 的 向量来表示它 ， 第一个维度表示它属于科幻电影的程度，第二个维度表示它属于浪漫电影的程度。我们可以预期的是，第一个维度的值会相当高，第二个维度的值则会相当低，然而一般也不会是 0 [[33]](#footnote-33)。

//构建数据，运行SVDPlusPlus  
val edges = spark.sparkContext.  
 makeRDD(*Array*(Edge(1L,11L,5.0),  
 Edge(1L,12L,4.0),  
 Edge(2L,12L,5.0),  
 Edge(2L,13L,5.0),  
 Edge(3L,11L,5.0),  
 Edge(3L,13L,2.0),  
 Edge(4L,11L,4.0),  
 Edge(4L,12L,4.0)))  
  
val conf = new lib.SVDPlusPlus.Conf(2,10,0,5,  
 0.007,0.007,0.005,0.015)  
val (g,mean) = lib.SVDPlusPlus.*run*(edges, conf)  
  
//函数pred()以及调用方法  
  
def pred(g:Graph[(Array[Double], Array[Double], Double, Double),Double],  
 mean:Double, u:Long, i:Long) = {  
 val user = g.vertices.filter(\_.\_1 == u).collect()(0).\_2  
 val item = g.vertices.filter(\_.\_1 == i).collect()(0).\_2  
 mean + user.\_3 + item.\_3 + item.\_1.zip(user.\_2).map(x => x.\_1 \* x.\_2).reduce(\_ + \_)  
}  
pred(g, mean, 4L, 13L)

## 关联规则

相关概念：

数据集：一般是购物篮数据

频繁模式：频繁地出现在数据集中的模式，列如项集，子结构和子序列等等

挖掘目标：频繁模式，频繁项集，关联规则

关联规则：比如牛奶 => 鸡蛋（支持度=2%，置信度-60%）

支持度：分析的数据集中的2%同时购买了牛奶和鸡蛋

支持度是两件商品（AB）在总销售笔数（N）中出现的概率，即A与B同时被购买的概率。

置信度：购买了牛奶的顾客有60%也购买了鸡蛋

置信度是购买A后再购买B的条件概率。简答来讲就是交集部分C在A中的比例，比例越大说明购买A的客户很大期望会购买B商品。

项集：商品或者品类的集合

k-项集：k个项组成的集合

频繁项集：满足最下支持度的项集，频繁k-项集一般记为Lk

强关联规则：满足最小支持度阈值和最小置信度阈值的规则

通过Apriori，FPGrowth算法挖掘多个保险产品之间的频繁项集。也就是，根据用户成单记录计算两两产品之间的相关度，比如：买了重大疾病类产品，推荐意外险，提高加保率。

获得的结果需要进行规则过滤：

互斥的商品，例如同类的自行车，汽车，内容相近的书籍，此时不能使用套餐推荐

推荐相近商品的时候使用浏览记录，使用购买记录更类似于关联规则挖掘

考虑兴趣的时效性，例如已经购买了某种自行车，就没必要再向用户推荐相近的自行车

总结一下：当我们推荐相近商品的时候，最好可以使用浏览记录来进行推荐，使用基于物品的协同过滤算法；

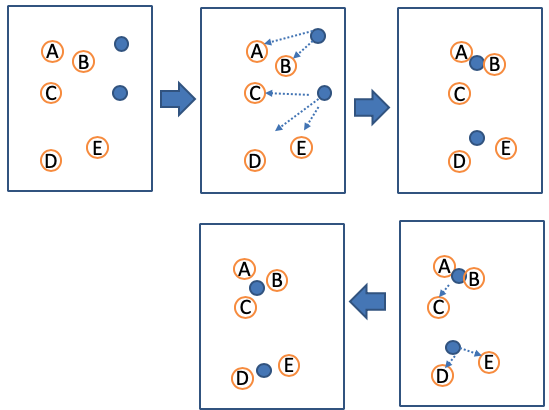
如果要推荐套餐的话，最好是基于购买记录，购物篮分析，用关联模式挖掘（关联规则挖掘）。

下面使用FP-Growth算法来找出高频推荐的电影。

该算的描述可在Han等人的论文“Mining frequent patterns without candidate generation”中找**2**到。FP代表frequent pattern，即高频模式。给定若干交易记录，FP-Growth的第一步是计算物品 的频率并标示高频物品。第二步是利用后缀树(suffix tree，也称为FP-tree)来编码各交易;该 过程不会显式生成推荐的备选集合，在大数据集上对应的计算量通常很大。

## K-means聚类算法

聚类算法在保险客户分群中起到了重要作用，K-means聚类算法首先会随机确定k个中心位置，位于空间中代表聚类中心的点，然后将各个数据项分配给最临近的中心点。待分配完之后，聚类中心就会移到分配给该聚类的所有节点的平均位置处，然后整个分配过程重新开始。这一过程会一直重复下去，直到分配过程不再产生变化位置。



上图模拟k=2的情况下如何进行聚类，第一个框中，两个中心点的位置是随机选择的，第二个框中把每个数据项分配给了距离最近的中心点，下面的框中进一步分配，最终结果是A、B、C在一个聚类中，D、E在另一个聚类中。

# 基于内容的推荐

## 标签体系的构建

用户和物品画像

图片包含 屏幕截图

描述已自动生成

标签系统和画像体系密不可分，构建画像绝大部分工作就是在计算标签。通过标签分类，包括事实标签和预测标签。事实标签。既定事实，从原始数据中提取。比如通过用户设置获取性别，通过实名认证获取生日，星座等信息。

模型标签。没有对应数据，需要定义规则，建立模型来计算得出标签实例。比如支付偏好度。预测标签。参考已有事实数据，来预测用户的行为或偏好。比如用户a的历史购物行为与群体A相似，使用协同过滤算法，预测用户a也会喜欢某件物品。

存储方式与结构：

采用文本数据库MongoDB和redis键值对数据库。MongoDB内存数据的应用主要在于对于单个用户的实时的查询，也是通过对spark数据梳理后的标签宽表进行数据格式转换(json格式)导入mongodb,前台应用可通过连接mongodb进行数据转换，从而进行单个标签的展现。(当然也可将数据转换为Redis中的key value形式，导入Redis集群)

## 构建用户画像

用户画像

用户画像是一个标签化的用户模型，它和推荐系统关系非常密切，一般意义上讲用户和物品的各种特征都可以通过概念抽象成为标签。

用户画像用于描述用户的基础属性、行为习惯和联系属性等等信心，对于业务了解用户具有非常重要的意义，可以帮助大幅度提升推荐的准确度。所以在搭建推荐的同事，一定要重视用户画像的建设。

**标签库**

标签是联系用户与物品、内容以及物品、内容之间的纽带，也是反应用户兴趣的重要数据源。标签库的最终用途在于对用户进行行为、属性标记。是将其他实体转换为计算机可以理解的语言关键的一步。

标签库则是对标签进行聚合的系统，包括对标签的管理、更新等。

一般来说，标签是以层级的形式组织的。可以有一级维度、二级维度等。

标签的来源主要有：

* 已有内容的标签
* 网络抓取流行标签
* 对运营的内容进行关键词提取

对于内容的关键词提取，使用[结巴分词](https://github.com/fxsjy/jieba) + [TFIDF](http://www.ruanyifeng.com/blog/2013/03/tf-idf.html)即可。此外，也可以使用[TextRank](http://www.tuicool.com/articles/UZ77Z3)来提取内容关键词。

这里需要注意的一点是对于关联标签的处理，比如用户的标签是足球，而内容的标签是德甲、英超，那么用户和内容是无法联系在一起的。最简单的方式是人工设置关联标签，此外也可以使用word2vec一类工具对标签做聚类处理，构建主题模型，将德甲、英超聚类到足球下面。

给用户推荐标签，可以提高标签的质量，并且方便用户输入标签。

## 构建物品画像

**基于物品标签的推荐系统**

豆瓣利用标签将用户的推荐系统结果做了聚类，显示了对不同标签下用户的推荐结果，从而增加了推荐的多样性和可解释性。

简单算法：

1. 统计每个用户最常用的标签、

2. 对于每个标签，统计被打过这个标签次数最多的物品。

3. 找到他常用的标签，然后找到具有这个标签的最热门物品推荐给这个用户。

利用用户的标签向量对用户兴趣建模，其中 每个标签都是用户使用过的标签，而标签的权重是用户使用该标签的次数。

可以对热门标签 和 热门物品进行惩罚，提高推荐物品的新颖性。

### 物品标签计算

**基于用户标签—群体的相似计算**

用户画像是目标用户的标签化表征，其中有些标签是定量的，有些标签是定性的。需要结合定量相似度和定性相似度来计算。

对不同的标签加权方法：

**定量相似度**

<user\_id, label\_id, label\_value> label\_id取标签，label\_value取实际值

将不同标签取值数据统一映射到[0,5]区间。常用归一化的方法有线性函数转换，数函数转换，反正切函数转换。

报错属性（交易属性）相似度

**定性相似度**

<user\_id, label\_id, label\_value> label\_id取二级标签类别，label\_value按照标签个数取值并归一化为[0-5]

包括属性（基础信息）相似度

所有的标签都属于定性相似度计算

**综合相似度**

最终的相似度为各标签的线性加权和。

**实验**

定量标签之间归一化到【0-5】

方法1： 定性相似度将标签对应二级类别下的Int类型，然后再归一化【0-5】

|  |  |  |  |
| --- | --- | --- | --- |
| user\_id | 用户ID |  | 归一化 |
| sex | 性别 | 1男， 2女， 0未知 | 归一化 |
| age | 年龄 |  | 归一化 |
| ID\_type | 证件类型 | 身份证，军官证， 护照， 台胞证，其它 | 不用 |
| ID\_number | 证件号码 | 注意ID是大写 | 不用 |
| birthday | 出生日期 |  | 不用 |
| lunar\_birthday | 出生日期(农历) |  | 不用 |
| occupation | 职业 | 学生， 公务员，教师，工程师，工人，农民，其它 | 不用 |
| education | 学历 | 小学， 初中， 高中/中专，大专/本科，研究生及以上 | 定性 |
| marital\_status | 婚姻 | 单身，结婚，离婚，丧偶，其它 | 定性 |
| income | 年收入 | 5万以下， 5万-10万， 10-20万， 20万以上 | 定性 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 年龄 | 110101 | 11010101 | 0-20岁 | 1 |
| 年龄 | 110101 | 11010102 | 21-30岁 | 2 |
| 年龄 | 110101 | 11010103 | 31-40岁 | 3 |
| 年龄 | 110101 | 11010104 | 41-50岁 | 4 |
| 年龄 | 110101 | 11010105 | 51-60岁 | 5 |
| 年龄 | 110101 | 11010106 | 61岁以上 | 6 |

方法2： 所有的标签的打分都为1，

比如<user\_id, 莘莘学子label\_id, 1> => <0101, 11010501,1>

### 商品评价数据的情感分析

## 利用标签进行推荐

## 基于时间和位置的推荐

1. 时间协同过滤[[34]](#footnote-34)

时间衰减模型，常用的衰减函数是一个指数函数

衰减率是自定义参数，用于调节时间的重要性。较大的表示时间越久，评分的重要性越低。

在预测阶段，时间下用户u的物品j的评分：

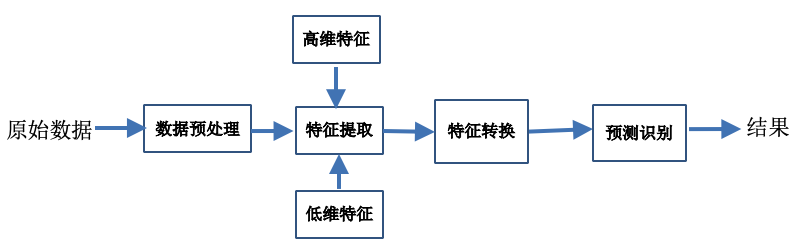
这里代表距离用户u对物品j的评分最近的k近邻。上述公式和传统的协同过滤最本质的区别在于预测函数中存在权重。这些权重会减低很久之前产生的评分，最终推荐方案也就更偏向最近的评分物品。

1. 位置感知推荐系统

位置感知推荐系统可以视为上下文感知推荐系统的特殊情况。

# 基于CTR预估的推荐

经典的推荐算法在解决推荐问题上解释力比较强，比如根据User-Based推荐解释可以说“与您相似的客户90%也买了这个商品”。但是，经典推荐算也存在很多问题，比如大数据量计算即为耗时，无法做到实时反馈。另外，每次计算都需要计算全量数据，学法做到增量学习和即时预测。基于点击率预估（CTR）的推荐算法是将推荐问题转化成分类或者排序问题。CTR预估不仅可以预测哪些用户会点击或者购买商品，也可以将候选池的商品进行排序并形成推荐列表。



## Spark分类算法[[35]](#footnote-35)

Spark中常见的3种分类模型:线性模型、决策树和朴素贝叶斯模型。线性模型 相对简单，而且相对容易扩展到非常大的数据集;决策树是一种强大的非线性技术，训练过程计 算量大并且较难扩展(幸运的是，MLlib会替我们考虑扩展性的问题)，但是在很多情况下性能 很好;朴素贝叶斯模型简单、易训练，并且具有高效和并行的优点(实际中，模型训练只需要遍 历整个数据集一次)。当采用合适的特征工程时，这些模型在很多应用中都能达到不错的性能。朴素贝叶斯模型还可以作为一个很好的模型测试基准，用于度量其他模型的性能。

目前，Spark的MLlib库提供了基于线性模型、决策树和朴素贝叶斯的二分类模型，以及基 于决策树和朴素贝叶斯的多类别分类模型。本书为了方便起见，将关注二分类问题。

### 线性模型

线性模型的核心思想是对样本的预测结果(通常称为目标变量或者因变量)进行建模，即对输入变量(特征或者自变量)应用简单的线性预测函数。

其中y是目标变量，w是参数向量(也称为权重向量)，x是输入特征向量。(wTx) 是关于权重向量w和特征向量x的线性预测器(又称向量点积)。对这个线性预测器，我们应用了一个函数f (又称连接函数)。

实际上，通过简单改变连接函数f，线性模型不仅可以用于分类还可以用于回归。标准的线 性回归(见下一章)使用对等连接函数(identity link，即直接使用y = fwT x )，而二分类使用上 面提到的连接函数。

举一个在线广告的例子。如果网页中展示的广告没有被点击，则目标变量标记为0；如果发生点击，则目标变量标记为1。每次点击的特征向量有点击事件的变量组成（比如与用户、网页、广告和广告客户相关的特征，以及与事件场景相关的其他因素，比如设备类型、事件、地理位置等等）。

于是，我们需要训练一个模型，将给定输入的特征向量映射到预测的输出（点击或者未点击）。对于一个新的数据点，我们将得到一个新的特征向量（此时还不知道预测的目标变量），并将其与权重向量进行点积。然后对点积的结果应用连接函数，最后函数的结果便是预测的输出。

给定输入数据的特征向量和目标变量，我们想要找到能够对数据进行最佳拟合的权重向量，拟合的过程即最小化模型输出与实际值的误差。这个过程称为模型的拟合、训练或者优化。具体来说，我们需要找到一个权重向量，它能够最小化所有训练样本的由损失函数计算出来的误差之和。损失函数的输入是给定训练样本的权重向量、特征向量和实际输出，而输出是误差值。实际上。损失函数也被定义为连接函数，每个分类或回归函数会有对应的损失函数。

MLlib提供的两个二分类模型的损失函数，第一个是logistic损失（logistic loss）等价于logistic回归模型。第二个是合页损失（hinge loss），等价于线性支持向量机（SVM）。这里的SVM严格来说并不属于广义线性模型的统计框架，但是当指定损失函数和连接函数时使用方法上相同。

1. Logistic回归

Logistic回归是一个概率模型，也就说该模型预测结果的值域为[0,1]。对于二分类来说logistic回归的输出等价于模型预测某个数据点属于正类的概率估计。Logistic回归是线性分类模型中使用最为广泛的一个。Logistic回归使用的连接函数为logistic连接：

Logistic回归的损失函数是logistic损失：

其中y是实际的输出值（正类为1，负类为-1）。

1. 线性支持向量机

SVM在回归和分类方面是一种强大且流行的技术。和logistic回归不同，SVM并不是概率模型，但是可以基于模型对正负的估计预测类别。

SVM的连接函数时一个对等连接函数，因此预测的输出为：

因此，当的估计值大于等于阈值0时，SVM将数据点标记为1，否则标记为0。

SVM的损失函数被称为合页损失，定义为：

SVM是一个最大间隔分类器，它试图训练一个使得类别尽可能分开的权重向量。在很多分类任务中，SVM不仅性能突出，而且在大数据集上的扩展线性的。

### 朴素贝叶斯

朴素贝叶斯是一个概率模型，通过计算给定数据点属于某个类别的概率来进行预测。朴素贝叶斯模型假定各个特征之间对分类的影响相互独立(假定各个特征之间条件独立)。

基于这个假设，属于某个类别的概率表示为若干概率乘积的函数，其中这些概率包括某个特 征在给定某个类别的条件下出现的概率(条件概率)，以及该类别的概率(先验概率)。这样使得 模型训练非常直接且易于处理。类别的先验概率和特征的条件概率可以通过数据的频率估计得 到。分类过程就是在给定特征和类别概率的情况下选择最可能的类别。

另外还有一个关于特征分布的假设，即参数的估计来自数据。MLlib实现了多项朴素贝叶斯 (multinomial naive Bayes)，其中假设特征分布是多项分布，用以表示特征的非负频率统计。

上述假设非常适合二元特征(比如k之一，k维特征向量中只有1维为1，其他为0)，并且 普遍用于文本分类(第4章中介绍的词袋模型是一个典型的二元特征表示)。

#### Spark-MLlib使用Bayes实现文本多分类

文本分类是指将一篇文章归到事先定义好的某一类或者某几类，在数据平台的一个典型的应用场景是，通过爬取用户浏览过的页面内容，识别出用户的浏览偏好，从而丰富该用户的画像。

分好词后，每一个词都作为一个特征，但需要将中文词语转换成Double型来表示，通常使用该词语的TF-IDF值作为特征值.

比如训练语料：1.txt

0,苹果 官网 苹果 宣布

1,苹果 梨 香蕉

逗号分隔的第一列为分类编号，0为科技，1为水果

import spark.implicits.\_  
  
将原始数据映射到DataFrame中，字段category为分类编号，字段text为分好的词，以空格分隔  
case class RawDataRecord(category: String, text: String)  
var dataDS = spark.read.textFile("/data/sougou-train").map {  
 x =>  
 var data = x.split(",")  
 *RawDataRecord*(data(0),data(1))  
}  
dataDS.select("category", "text").show(2)  
  
80%作为训练数据，20%作为测试数据  
val splits = dataDS.randomSplit(*Array*(0.7, 0.3))  
var trainingDF = splits(0).toDF()  
var testDF = splits(1).toDF()  
  
//将分好的词语转换成数组  
var tokenizer = new Tokenizer().setInputCol("text").setOutputCol("words")  
var wordsData = tokenizer.transform(trainingDF)  
*println*("output1：")  
wordsData.select($"category",$"text",$"words").take(1)

//计算每个词在文档中的词频  
var hashingTF = new HashingTF().  
 setNumFeatures(500000).  
 setInputCol("words").  
 setOutputCol("rawFeatures")  
var featurizedData = hashingTF.transform(wordsData)  
*println*("output2：")  
featurizedData.select($"category", $"words", $"rawFeatures").take(2)

这里将中文词语转换成INT型的Hashing算法，类似于Bloomfilter，上面的setNumFeatures(100)表示将Hash分桶的数量设置为100个，这个值默认为2的20次方，即1048576，可以根据你的词语数量来调整，一般来说，这个值越大，不同的词被计算为一个Hash值的概率就越小，数据也更准确，但需要消耗更大的内存，和Bloomfilter是一个道理[[36]](#footnote-36)。

每个分类下有几千个文档，这里将这些语料进行分词，然后每一个分类生成一个文件，在该文件中，每一行数据表示一个文档的分词结果，重新用0-9作为这10个分类的编号：

0 汽车

1 财经

2 IT

3 健康

4 体育

5 旅游

6 教育

7 招聘

8 文化

9 军事

### 决策树

决策树是一种强大的非概率模型，它可以表达复杂的非线性模式和特征相互关系。决策树在 很多任务上表现出的性能很好，相对容易理解和解释，可以处理类别特征和数值特征，同时不要 求输入数据归一化或者标准化。决策树非常适合应用集成方法，比如多个决策树的集成(称为决策树森林)。

决策树模型就好比一棵树，叶子代表值为0或1的分类，树枝代表特征。下图展示了一棵简 单的决策树，二元输出分别是“待在家里”和“去海滩”，特征则是天气。

决策树算法是一种自上而下的、始于根节点（或特征）的方法，在每一个步骤中通过评估特征分割的信息增益，选出分割数据集最优的特征。信息增益通过计算节点不纯度（节点标签不相似或不同质的程度）减去分割后的两个子节点不纯度的加权和。对于分类任务，有两种评估方法可用于选择最好的分割：基尼不纯度（Gini impurity）和熵（entropy）。

### 树集成模型

集成模型指将基础模型组合成一个模型。Spark支持两种主要的集成算法：随机森林和梯度提升树（GBDT）。

1. 随机森林

随机森林即决策树的集成，它由多个决策树组合而成。如决策树一样，随机森林能处理类别特征、支持多分类而且不需要特征缩放。

Spark MLlib的随机森林算法同时支持二分类和多类别分类，以及连续型和类别型特征上的 回归。

1. 梯度提升树

梯度提升树是决策树的集成。它迭代地对决策树进行训练以最小化损失函数。它能处理类别性特征、支持多类别分类且不需要特征缩放。

3. 多层感知分类器

神经网络是一个复杂的自适应系统，它会借助各权重的变更而改变信息流，进而改变自己的 内部结构。针对多层神经网络的权重优化过程也称为反向传播(backpropagation)。反向传播超出 了本书讨论范围，另也涉及激活函数和基本的微积分知识。

多层感知分类器(multilayer perceptron classifier)基于前向反馈(feed-forward)人工神经网6络。它由多个神经层构成，每层都与下一层全连接。其输入层的各节点对应输入数据。其他节点 都会对经该节点的输入、相应的权重和偏置(bias)进行线性组合，再应用一个激活函数(activation function)或连接函数后，映射为对应的输出。

## GBDT+LR实现高潜客户识别和Item排序

客户特征的二分类建模，训练数据包括所有已电销的通话记录。预测新客户可能购买保险的概率，选择购买意向高的用户为最终客户。

图片包含 标牌, 文字

描述已自动生成

该模型的实现比较注重客户的分组情况，需要提前通过聚类的方式计算潜在的分析特征，比如客户可能同属于同一种保险购买群体。训练数据特征化之后，结合树形结构的集成算法GBDT得到多个基模型的结果，这里我们不会通过残差的加减来减少偏差，而是加入LR线性模型得到每个分类树的概率。

基于Spark的GBDT+LR模型实现测试数据来源<http://archive.ics.uci.edu/ml/machine-learning-databases/adult/>[[37]](#footnote-37)

该模型利用Spark mllib的GradientBoostedTrees作为GBDT部分，因为ml模块的GBTClassifier对所生成的模型做了相当严密的封装，导致难以获取某些类或方法。而GradientBoostedTrees所需的训练数据为mllib下的LabeledPoint，所以下面的数据预处理的目标是将cat数据进行编码并生成LabeledPoint。

数据预处理

GBDT部分

GBDT+LR

### xgboost+LR

背景[[38]](#footnote-38)

xgboost+lr模型融合方法用于分类或者回归的思想最早由facebook在广告ctr预测中提出，其论文Practical Lessons from Predicting Clicks on Ads at Facebook有对其进行阐述。在这篇论文中他们提出了一种将xgboost作为feature transform的方法。大概的思想可以描述为如下：先用已有特征训练XGBoost模型，然后利用XGBoost模型学习到的树来构造新特征，最后把这些新特征加入原有特征一起训练模型。构造的新特征向量是取值0/1的，向量的每个元素对应于XGBoost模型中树的叶子结点。当一个样本点通过某棵树最终落在这棵树的一个叶子结点上，那么在新特征向量中这个叶子结点对应的元素值为1，而这棵树的其他叶子结点对应的元素值为0。新特征向量的长度等于XGBoost模型里所有树包含的叶子结点数之和。最后将新的特征扔到LR模型进行训练。实验结果表明xgboost+lr能取得比单独使用两个模型都好的效果。

**原理介绍**

“特征决定了所有算法效果的上限，而不同的算法只是离这个上限的距离不同而已。”所以如何更有效的提取有效的特征是机器学习中的一个hotspot，例如近几年来大火的深度学习方法中神经网络的层数不断增加其实质也是在探索如何更好地从原始数据中得到更为有效的特征表达。如果能够将数据表达成为线性可分的数据，那么使用简单的线性模型就可以取得很好的效果。

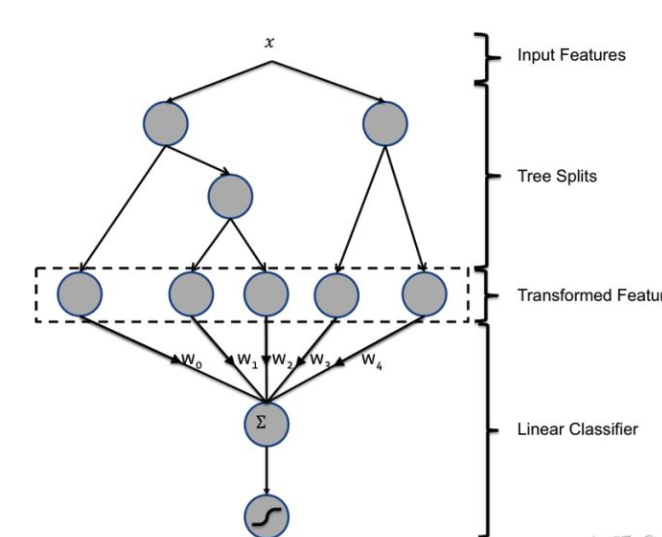
XGBoost构建新的特征也是使特征更好地表达数据。它的核心思想是将boosting看作是一个将样本进行非线性变换的方法。处理特征变换的一般方法有：

对于连续的特征：一个简单的非线性变化就是将特征划分到不同的区域(bin)，然后再将这些区域的编号看作一个离散的特征来进行训练。这也就是俗称的连续变量离散化方法，这有非常多的方法可以完成这项事情。

对于离散的特征：我们可以直接对离散特征做一个笛卡尔积从而得到一系列特征的组合，当然有些组合是没用的，那些没用的组合可以删掉。

而这里利用boosting来对样本进行离散化的方法是另一种特征变换的方法。

下面的图中的两棵树是利用现有特征训练XGBoost学习到的，其中第一棵树有3个叶子结点，而第二棵树有2个叶子节点。对于一个输入样本点x，如果它在第一棵树最后落在其中的第二个叶子结点，而在第二棵树里最后落在其中的第一个叶子结点。那么通过XGBoost获得的新特征向量为[0, 1, 0, 1, 0]，其中向量中的前三位对应第一棵树的3个叶子结点，后两位对应第二棵树的2个叶子结点。下图为混合模型结构。输入特征通过增强的决策树进行转换。 每个单独树的输出被视为稀疏线性分类器的分类输入特征。 增强的决策树被证明是非常强大的特征转换。



这里我们可以这样理解，XGBoost是一种集成树模型，其本质是不同单个决策树的组合。而决策树是一种树形结构，又称为判定树，是运用于分类的一种树结构，其中的每个内部节点代表对某一属性的一次测试，每条边代表一个测试结果，叶节点代表某个类或类的分布。决策树的决策过程需要从决策树的根节点开始，待测数据与决策树中的特征节点进行比较，并按照比较结果选择选择下一比较分支，直到叶子节点作为最终的决策结果。

未完待续。。。

### xgboost4j-spark[[39]](#footnote-39)

xgboost4j-spark使得xgboost能够集成spark项目，运行在分布式集群中。

当前的xgboost4j-spark可以实现以下功能：

特征工程：特征提取，变换，降维和选择等。 管道：构造，评估和调整ML管道 持久性：持久并加载机器学习模型，甚至整个管道

如何使用XGBoost4J-Spark构建机器学习管道的端到端过程，需要了解以下内容： 使用Spark预处理数据以适合XGBoost / XGBoost4J-Spark的数据接口 使用XGBoost4J-Spark训练XGBoost模型 使用Spark服务XGBoost模型（预测） 使用XGBoost4J-Spark构建机器学习管道 在生产中运行XGBoost4J-Spark

添加Xgboost4j-Spark依赖包，如果你使用的是Maven进行项目管理，添加

<dependency>

<groupId>ml.dmlc</groupId>

<artifactId>xgboost4j-spark</artifactId>

<version>latest\_version\_num</version>

</dependency>

注：lastest\_version\_num是当前具体的版本号，详细见<https://github.com/dmlc/xgboost/releases>

Xgboost4j-spark需要Spark的版本大于2.4

因为xgboost4j-spark高度集成了xgboost和spark，使用户可以通过便捷而强大的数据处理框架Spark在训练/测试数据集上应用各种类型的转换。下面以鸢尾花(iris)数据集为例，展示如何使用Spark转换原始数据集并使之适合XGBoost的数据接口。

鸢尾花数据集是csv格式。每个实例包含4个特征，即“sepal lentgh”，“septal width”，“petal length”和“petal length”。此外，它包含“ class”列，该列实际上是带有三个可能值的标签：“ Iris Setosa”，“ Iris Versicolour”和“ Iris Virginica”。

读取数据集，并转换成DataFrame：

import org.apache.spark.sql.SparkSession  
import org.apache.spark.sql.types.{DoubleType, StringType, StructField, StructType}  
  
val spark = SparkSession.builder().getOrCreate()  
val schema = new StructType(Array(  
 StructField(“sepal length”, DoubleType, true),  
 StructField(“sepal width”, DoubleType, true),  
 StructField(“petal length”, DoubleType, true),  
 StructField(“petal width”, DoubleType, true),  
 StructField(“class”, StringType, true)))  
val rawInput = spark.read.schema(schema).csv(“input\_path”)

创建SparkSession作为DataFrame的入口，schema变量定义鸢尾花数据集的数据格式，这里通过csv读取文件。

为了适应xgboost，需要对原始iris数据集进行转换，比如String类型的列转换成Double类型，还有要将所有特征列组合成向量。

String类型转Double类型，可以使用Spark内置的StringIndexer类。

import org.apache.spark.ml.feature.StringIndexer  
val stringIndexer = new StringIndexer().  
 setInputCol(“class”).  
 setOutputCol(“classIndex”).  
 fit(rawInput)  
val labelTransformed = stringIndexer.transform(rawInput).drop(“class”)

如上，创建一个StringIndexer实例，设置String类型作为输入列，设置Double类型作为输出类型。然后使用fit将rawInput作为输入DataFrame。

定义好StringIndexer的转换逻辑之后，使用transform输入rawInput，会生成一个新的列ClassIndex，这里删除掉Class列。

类似的，也可以使用另外一种转换操作VectorAssembler，这个操作可以将多个特征组合在一起形成向量。

import org.apache.spark.ml.feature.VectorAssembler  
val vectorAssembler = new VectorAssembler().  
 setInputCols(*Array*(“sepal length”, “sepal width”, “petal length”, “petal width”)).  
 setOutputCol(“features”)  
val xgbInput = vectorAssembler.transform(labelTransformed).select(“features”, “classIndex”)

现在xgbInput这个DataFrame之后两列，分别是features和classIndex。前者是特征向量，后者是数值标签列。

处理缺失值

训练  
val *Array*(train, eval1, eval2, test) = xgbInput.randomSplit(*Array*(0.6, 0.2, 0.1, 0.1))

val xgbParam = *Map*(“eta” -> 0.1f,  
 “max\_depth” -> 2,  
 “objective” -> “multi:softprob”,  
 “num\_class” -> 3,  
 “num\_round” -> 100,  
 “num\_workers” -> 2,  
 “eval\_sets” -> *Map*(“eval1” -> eval1, “eval2” -> eval2))  
val xgbClassifier = new XGBoostClassifier(xgbParam).  
 setFeaturesCol(“features”).  
 setLabelCol(“classIndex”)  
val xgbClassificationModel = xgbClassifier.fit(train)

预测

val results = xgbClassificationModel.transform(test)  
results.show()

处理缺失值

构建pipeline

未完待续。。。

## FM因式分解机

因式分解机（Factorization Machine）是一种基于矩阵分解的机器学习模型，可以轻松因对高度稀疏的数据场景。FM允许更多的特征工程，可以将所有的特征表示成embedding vector，构造二阶关系，同时，在联合特征的权重捕捉上，不再采用简单的单权重，而是采用向量内积来表示，这样就能从原始信息中捕捉更多的信息。FM可以再线性时间复杂度内完成非线性的计算任务，有利于拟合训练集中的非线性关系。

图片包含 物体

描述已自动生成

如上损失函数，最后加入了组合特征。一般线性模型在处理组合特征时泛化能力弱，但FM增加了xi,xj这种组合特征，并且vi,vj就是特征学习之后的embedding向量的内积，这就可以代表权重。注重有效的权重赋值方法增加了FM的泛化能力较强。

在2014年两个CTR预测比赛中获胜的方法都使用了一种叫Field-aware Factorization Machines (FFM)的方法，它是因式分解机(Factorization Machines)的变体。FFM尝试通过学习每个特征交互对的潜在因素来为特征交互建模。这个算法可以在LibFFM框架中实现并且已被许多参赛者使用。LibFFM对大型数据集的并行处理和内存使用非常有效。

## 强化学习 - Clustering Of Bandit

Clustering Of Bandit简称CLUB，其在推荐系统领域主要解决EE问题，尤其对冷启动阶段效果较好。我们假设，用户成交了一份保单，其反馈回报和用户相关的feature成线性关系。由此，用客户和产品的特征预估回报和其置信区间，然后选择置信区间上界最大的产品作为推荐结果。

图片包含 屏幕截图

描述已自动生成

CLUB跟其它Bandit算法最大的区别在于它引入了用户特征的计算，让整个算法收敛更快。

## 排序学习（Learning To Rank）

排序是对一组物品列表按照某种方式进行排序，来最大化整个列表的效用的过程，广泛应用于搜索引擎、推荐系统、机器翻译、对话系统甚至计算生物学。一些监督机器学习技术经常被广泛应用在这些问题中，这些技术称作排序学习技术。

排序学习(Learning to Rank, LTR)最早兴起于信息检索领域。经典的信息检索模型包括布尔模型、向量空间模型 、 概率模型、语言模型以及链接分析等。这些在不同时期提出的模型都是无监督排序方法，其共同特点是利用一些简单的特征进行排序，例如词频、逆文档频率等。这些传统排序方法的优点在于容易进行经验参数的调整，得到最优的参数，用以对检索文档按照一定标准(往往是查询与文档之间的相关性)进行排序。

随着搜索引擎需要处理的数据量呈指数增长，人为凭经验优化参数的过程变得越来越复杂。另外，影响文档排序的信息多种多样（文本相关性中的VSM、LM、BM25分值；临近度；PageRank值；是否是垃圾网页；URL深度；域名重要性等），然后这些经典的模型往往偏重某一方面的因素而忽略了其他可以用于排序的重要因素，例如概率模型和语言模型都没有考虑网页链接、网页pagerank值等互联网结构对排序的影响。在这种情况下，排序学习应运而生。排序学习的定义为：基于机器学习中用于解决分类与回归问题的思想，提出利用机器学习方法解决排序的问题。排序学习的目标在于自动地从训练数据中学习得到一个排序函数，使其在文本检索中能够针对文本的相关性、重要性等衡量标准对文本进行排序。机器学习的优势是：整合大量复杂特征并自动进行参数调整，自动学习最优参数，降低了单一考虑排序因素的风险；同时，能够通过众多有效手段规避过拟合问题。

而在推荐系统领域，传统的推荐算法主要可以分为3大类：基于内容的推荐算法、协同过滤推荐算法以及混合推荐算法。这些传统推荐算法重点考虑用户和物品之间的二元关系，大都可以转化为评分预测问题，根据用户对物品的评分进行排序后产生推荐列表。但仅仅依据用户对物品的评分产生推荐结果并不能准确地体现用户的偏好。随着数据的增长，关于用户、物品、上下文、搜索等信息的获取成为了可能，而这些因素对用户的偏好有很大的影响。单一的推荐模型仅能从某一个方面的因素考虑用户的偏好，此时就需要排序模型来综合考虑各种影响用户偏好的因素，推荐系统领域的排序学习也因此应运而生[[40]](#footnote-40)。

Learning To Rank(LTR)算法有三种类型，分别是pointwise、pairwise和listwise排序。

在协同过滤算法中，推荐本身可以视作一个评分预测系统。用户-物品的预测评分越高，该用户就越可能喜欢该物品。

而在实际推荐场景中，用户能够看到的页面中并不会显示评分，而是各个物品的排序列表。也就是说，推荐结果关注了用户对于物品的相关性而非评分。

PointWise

PointWise类似于协同过滤的方法，先为每个项目进行评分，然后在按照评分进行排序。评分预测与排序的区别在于，用户不用关心物品的实际评分，只要能够体现物品的排序位置就行。

PairWise

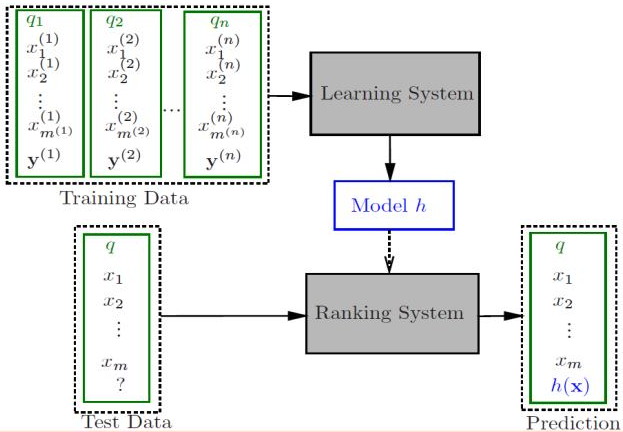
PairWise是二分类的一种方法。它输入两个物品，返回两个物品的顺序。如果得到所有物品两两组合的顺序，最终就能确定所有物品的排序结果。

该方法与PointWise的本质不同在于，pairwise并不需要预测所有商品的评分。 ListWise Listwise是所有LTR中终极方法，它根据整个排名列表并进行优化。listwise排序基于这样一个假设，物品排序列表顶部的顺序比底部的顺序更重要。这也是listwise的一大优势，因为pointwise和pairwise都不能区分某个物品在排名列表中的位置。

比如，一共推荐Top10的商品。pairwise推荐过程中，排名靠前的两个商品的顺序与排名最后的两个商品的位置重要度一样。从直接上将，这显然不符合实际情况，众所周知，用户更多的只会关心列表顶部的商品。基于这一角度，listwise的优势非常明显。

在推荐系统领域，根据业务数据可以划分为显式(Explicit)和隐式(Implicit)反馈类型。基于近邻(neighbourhood-based)和矩阵分解(MF)的方法预测显式反馈的评分效果更好，而对于隐式数据效果一般。针对评分数据设计的矩阵分解算法，类似于监督学习中的回归问题，而隐式反馈数据通常是0-1二值数据。

每个Item的打分 +… 根据已有的成单训练数据作为标记对象，根据分类或回归的方式获得a,b,c,d参数的最优组合，然后通过这个打分函数进行相关性判断。



对于隐式反馈的数据，我们需要换一个思路：“预测用户是否会对Item产生某种隐式反馈”的分类问题。推荐列表注重排序，并不关心单个item预测的评分多少。Pair-wise Raiking就是对同一个User，直接考虑任意两个Item的顺序谁在前谁在后。

根据以上思路，把排序问题转换为一个分类问题。分类问题的训练和预测需要正负样本，但是，通常情况下隐式反馈数据的负样本不那么明显。比如，很难肯定的认为用户没有点击或者是买过的商品一定是用户不感兴趣的。因为有可能是用户没有看到或者是暂时没有兴趣，但是不代表以后不会产生兴趣，

这时就需要借助排序学习，排序问题不会考虑正负样本的划分，而是计算样本的前后顺序。推荐列表注重排序，并不关心单个item预测的评分多少。

下面介绍两个排序学习的主流方法：

planA： Point Wise类：还是预测User对单个Item的偏好，最后根据偏好排序。

planB： Pair Wise类：对同一个User，直接考虑任意两个Item的顺序谁在前谁在后。

Pointwise和Pairwise的原理分析：

Pointwise方法的主要思想是将排序问题转化为多类分类问题或者回归问题。

Pairwise的主要思想是将Ranking问题形式化为二元分类问题。

第一，pointwise解决的是单文档分类问题，模型会告诉我们该条文档属于good，fair，bad其中一类，模型根据给出的训练数据进行建模，那么训练的数据我们首先只有搜出的文档，然后开始人工标注，或者根据用户反馈等等一系列手段，标注会根据一次搜索的返回结果考虑每个文档的好坏程度，但是不会比较两次不同搜索的返回结果好坏，我觉得这个就是截图中要说明的问题。所以，在这种情况下，模型可能会出现问题。

第二，pairwise解决的是文档对的排序问题，其实归根结底也是个分类问题，文档1排在文档2的几率问题，pairwise会考虑相同搜索的情况下，返回结果的排序问题，但是有一个弊端就是如果中间有一次预测失误，那么对余下数据的根据错误的数据进行比较，就会出现问题。

## TensorFlow构建DNC模型

数据采用criteo display ads[[41]](#footnote-41)

未完待续。。。

## 算法总结

综上所述，Memory\_Based算法和监督学习算法存在很多差异，从模型角度来说，Memory\_Based算法不需要生成模型来预测客户-物品之间匹配的概率，而是通过计算相似度或者隐因子关系，最终生成推荐结果；而监督学习算法需要通过训练集学习生成模型，再通过预测集的特征向量来预测。

从业务角度看，Memory\_Based算法适合于显式反馈的业务场景，比如豆瓣电影和亚马逊图书，这些场景需要用户给自己接触的物品打分，而打分就直接对应用户的喜好度。但是在其他业务场景，比如金融和保险产品中，用户购买产品并不能直接得到相应的喜好度和评分，用户可能是为了别人买保险，或者只是生活需要。我们把后一种场景得到数据为隐式反馈数据。所以，通过隐式反馈数据最终只能推荐用户适合的物品，而非用户最终喜爱的物品。

# 推荐系统冷启动

冷启动问题是关系新用户的推荐问题。如何给用户分群，以此来进行半个性化推荐。学习一个用关联规则实现的简单个性化推荐。

如果推荐系统对于一个用户的相关信息一无所知，很显然是无法进行个性化推荐的。但是推荐系统的目的就是通过个性化推送取悦用户，这就让我们陷入两难。而这个难题就是冷启动问题。实际上，冷启动问题不仅仅是针对于用户而言，刚上架的物品也存在相同的问题，你没有办法确定这个物品该推荐给哪些用户。所以，新物品通常不会出现在用户的个性化推荐中。

不过正常情况下，对于真正的冷启动用户，他没有关于任何物品的相关行为（包括点击、浏览、购买等等），这种用户数量也是及其有限的。对于另外一种情况，比如某些用户的品味比较独特，与他相似的用户几乎没有，这也属于冷启动的范畴。

物品冷启动

比较好的解决方法是将新上架物品单独分类，让所有用户都有机会看到。比如添加一个新商品的按钮，点击进去就是最近的新商品列表。

不过这只限于产品线不大的推荐场景，如果是每日新上架商品众多，比如大型电商网站，这个方法就不奏效了。

使用算法解决冷启动问题

实际上，冷启动问题并没有一个很好的解决办法。即使是算法业务能力为了，算法本身也需要建立在数据之上。

使用关联规则给冷启动用户提供推荐

利用领域知识和业务规则进行推荐

用户分群

利用人口统计信息进行用户分群，然后进行分组推荐。

比较显著的用户分群

根据地理位置进行过滤。比如如果系统定位到用户的登录位置为上海，可以将购买地址为上海的商品过滤并推荐给他。

根据性别进行过滤。

用户在什么时候不在是冷启动用户？

1、用户冷启动

用户冷启动主要解决如何给新用户做个性化推荐的问题。当新用户到来时，我们没有他的行为数据，所以也无法根据他的历史行为预测其兴趣，从而无法借此给他做个性化推荐。

2. 物品的冷启动

将新物品推荐给曾经喜欢过与新物品相似的物品的用户，

由于新item没有用户行为，因此物品相似度只能从item自身出发，包括根据item的内容信息挖掘出item的向量表示，通过向量相似度来刻画物品相似度，还可以利用topic model挖掘出item的主题分布等等。

3. 系统冷启动

比较热门 --- 需要过滤掉比较冷门的物品，因为用户不知道背景信息，很难进行反馈

具有代表性和区分性—过滤大众化和老少皆宜的物品，因为用户的反馈无法体现个性化

启动物品集合需要有多样性—品类覆盖率要高，包含所有主流的用户兴趣

## EE问题

推荐系统领域有两个经典问题，EE问题和冷启动问题。EE即探索（exploration）和利用（exploitation）问题。“利用”就是要迎合用户的习惯，推荐用户感兴趣的东西；“探索”就是说不能只是推荐用户当前感兴趣的东西，久而久之，用户必然会腻烦，所以既要照顾客户当前兴趣又要探索用户新的兴趣，从评价角度讲就是平衡准确率和多样性。

冷启动问题设计到产品算法运营等一系列问题。Bandit算法是一种简单的在线学习算法，常常用于解决这个问题，下面介绍基础的Bandit算法以及相关的改进算法。

## Bandit算法

假设你走近一家赌场，面前有一排老虎机，外表一模一样，但是每个老虎机吐钱的概率可不一样。如何知道每个老虎机吐钱的概率分布，每次如何选择老虎机可以做到最大化收益呢？这就是多臂赌博机问题（Multi-armed bandit problem,MAB）。

如何去解决这个问题，最好的办法是试一试，不能盲目的试，而是有策略并且快速的试，这些策略就是Bandit算法[[42]](#footnote-42)。

未完待续。。。

# 推荐系统的评价方式

推荐系统的评价标准有很多。

用户满意度（User Satifaction），调研或用户反馈；点击率、转化率等。

准确率（Accuracy）：precision、recall、f-score

覆盖率（Coverage）：召回到尾部物品和用户

多样性（Diversity）：两两之间不相似

新颖性（Novelty）：用户不熟悉的物品

惊喜性（Serendipity）：如何评价？

用户信任度（Trust）/可解释性（Explanation）：推荐理由

鲁棒性/健壮性（Robustness）：哈利波特现象；抗攻击、反作弊

实时性（Real-time/Online）：新加入的物品；新的用户行为（实时意图）

商业目标（business target）：一个用户带来多少盈利

如何判断推荐的结果是否符合用户的需求？这就需要某种方式来评估结果的预测效果。评估指标就是那些衡量模型预测能力或准确率的方法。我们看下Spark如何计算推荐系统和协同过滤模块里面常用的两个指标：均方差（mean squared error）和K值平均准确率（mean average precision at K）[[43]](#footnote-43)。

## 评分预测

### 均方差

均方差(MSE，mean squared error)直接衡量“用户-物品”评级矩阵的重建误差。它也是 一些模型里所采用的最小化目标函数，特别是许多矩阵分解类方法，比如ALS。因此，它常用于显式评分的情形。

它的定义为各平方误差的和与总数目的商。其中平方误差是指预测到的评级与真实评级的差 值的平方。

下面以用户789为例来讲解。现在从之前计算的moviesForUser这个Ratings集合里找 出该用户的第一个评级:

val actualRating = moviesForUser.take(1)(0)

输出为:

actualRating: org.apache.spark.mllib.recommendation.Rating = Rating(789,1012,4.0)

可以看到该用户对该电影的评级为4。然后，求模型的预计评级: val predictedRating = model.predict(789, actualRating.product)

其输出是:

...

MatrixFactorizationModel.scala:46, took 0.025404 s

predictedRating: Double = 4.001005374200248

可以看出模型预测的评级差不多也是4，十分接近用户的实际评级。最后，我们计算实际评 级和预计评级的平方误差:

val squaredError = math.pow(predictedRating - actualRating.rating, 2.0)

这将输出:

squaredError: Double = 1.010777282523947E-6

要计算整个数据集上MSE，需要对每一条(user,item,actual rating, predicted rating)记录都计算该平均误差，然后求和，再除以总的评级次数。具体实现如下。

以下代码取自Apache Spark编程指南中的ALS部分:http://spark.apache.org/docs/latest/mllib-collaborative-filtering.html。

首先从ratings RDD里提取用户和物品的ID，并使用model.predict来对各个“用户-物品“对各个”用户-物品“对做预测。所得的RDD以“用户和物品ID”对作为主键，对应的预计评级作为值:

val usersProducts = ratings.map{ case Rating(user, product, rating) => (user, product)

}

val predictions = model.predict(usersProducts).map{

case Rating(user, product, rating) => ((user, product), rating)

}

接着提取出真实的评级。同时，对ratings RDD做映射以让“用户-物品”对为主键，实6际评级为对应的值。这样，就得到了两个主键组成相同的RDD。将两者连接起来，以创建一个 新的RDD。 这个RDD的主键为“用户-物品”对，键值为相应的实际评级和预计评级。

val ratingsAndPredictions = ratings.map{

case Rating(user, product, rating) => ((user, product), rating)

}.join(predictions)

最后，求上述MSE。具体先用reduce来对平方误差求和，然后再除以count函数所求得的总记录数：

val MSE = ratingsAndPredictions.map{

case ((user, product), (actual, predicted)) => math.pow((actual - predicted), 2)

}.reduce(\_ + \_) / ratingsAndPredictions.count

println(“Mean Squared Error = “ + MSE)

对应的输出如下:

Mean Squared Error = 0.08231947642632852

均方根误差(RMSE，root mean squared error)的使用也很普遍，其计算只需在MSE上取平 方根即可。这不难理解，因为两者背后使用的数据(即评级数据)相同。它等同于求预计评级和 实际评级的差值的标准差。如下代码便可求出:

val RMSE = math.sqrt(MSE)

println(“Root Mean Squared Error = “ + RMSE)

其输出的均方根误差为:

Root Mean Squared Error = 0.2869137090247319

结合该误差的定义来对上述结果进行理解。将RMSE误差降低，即使将预测值向理想值拟 合。另外也需注意实际数据的最大值和最小值。

### K值平均准确率

K值平均准确率(MAPK)的意思是整个数据集上的K值平均准确率(APK，average precision at K metric)的均值。APK是信息检索中常用的一个指标。它用于衡量针对某个查询所返回的“前K个”文档的平均相关性。对于每次查询，我们会将结果中的前K个与实际相关的文档进行比较。

用APK指标计算时，结果中文档的排名十分重要。如果结果中文档的实际相关性越高且排 名也更靠前，那APK分值也就越高。由此，它也很适合评估推荐的好坏，因为推荐系统也会计 算“前K个”推荐物，然后呈现给用户。如果在预测结果中得分更高(在推荐列表中排名也更靠 前)的物品实际上也与用户更相关，那自然这个模型就更好。APK和其他基于排名的指标同样 也更适合评估隐式数据集上的推荐。这里用MSE相对就不那么合适。

当用APK来做评估推荐模型时，每一个用户相当于一个查询，而每一个“前K个”推荐物 组成的集合则相当于一个查到的文档结果集合。用户对电影的实际评级便对应着文档的实际相关 性。这样，APK所试图衡量的是模型对用户感兴趣和会去接触的物品的预测能力。

实际场景中，可以分不同的数据流程单独进行评价，比如召回流程，用召回率评价召回商品的质量。

而在排序阶段可以用NDCG或MRR

分别计算MSE、RMSE、精准率、召回率、覆盖率、流行度、多样性

import math

def RMSE(records):

“””计算RMSE

@param records: 预测评价与真实评价记录的一个list

@return: RMSE

“””

numerator = sum([(pred\_rating - actual\_rating)\*\*2 for pred\_rating, actual\_rating in records])

denominator = float(len(records))

return math.sqrt(numerator / denominator)

def MSE(records):

“””计算MSE

@param records: 预测评价与真实评价记录的一个list

@return: MSE

“””

numerator = sum([(pred\_rating - actual\_rating)\*\*2 for pred\_rating, actual\_rating in records])

denominator = float(len(records))

return numerator / denominator

def precision(recommends, tests):

“””计算Precision

:param recommends: dict

给用户推荐的商品，recommends为一个dict，格式为 { userID : 推荐的物品 }

:param tests: dict

测试集，同样为一个dict，格式为 { userID : 实际发生事务的物品 }

:return: float

Precision

“””

n\_union = 0.

user\_sum = 0.

for user\_id, items in recommends.items():

recommend\_set = set(items)

test\_set = set(tests[user\_id])

n\_union += len(recommend\_set & test\_set)

user\_sum += len(test\_set)

return n\_union / user\_sum

def recall(recommends, tests):

“””

计算Recall

@param recommends: 给用户推荐的商品，recommends为一个dict，格式为 { userID : 推荐的物品 }

@param tests: 测试集，同样为一个dict，格式为 { userID : 实际发生事务的物品 }

@return: Recall

“””

n\_union = 0.

recommend\_sum = 0.

for user\_id, items in recommends.items():

recommend\_set = set(items)

test\_set = set(tests[user\_id])

n\_union += len(recommend\_set & test\_set)

recommend\_sum += len(recommend\_set)

return n\_union / recommend\_sum

def coverage(recommends, all\_items):

“””

计算覆盖率

@param recommends : dict形式 { userID : Items }

@param all\_items : 所有的items，为list或set类型

“””

recommend\_items = set()

for \_, items in recommends.items():

for item in items:

recommend\_items.add(item)

return len(recommend\_items) / len(all\_items)

def popularity(item\_popular, recommends):

“””计算流行度

@param item\_popular: 商品流行度　dict形式{ itemID : popularity}

@param recommends : dict形式 { userID : Items }

@return: 平均流行度

“””

popularity = 0. # 流行度

n = 0.

for \_, items in recommends.items():

for item in items:

popularity += math.log(1. + item\_popular.get(item, 0.))

n += 1

return popularity / n

## 分类的评价方式

ROC曲线：接受者操作特征曲线（receiver operating characteristic curve)，是反映名感性和特异性连续变量的综合指标，ROC曲线上每个点反映着对同一想好刺激的感受性。

AUC(Area Under Curve)是ROC曲线下的面积，显然这个面积的数值不会大于1。由于ROC曲线一般都处于y=x这条直线的上方，所以AUC的取值范围一般在0.5到1之间。使用AUC值作为评价标准是因为很多时候ROC曲线并不能清晰的说明哪个分类器的效果更好，而作为一个数值，对应AUC更大的分类器效果更好。

AUC的计算有两种方式，梯形法和ROC AUCH法，都是以逼近法求近似值。

AUC值是一个概率值。当你随机挑选一个正样本和一个负样本，当前的分类算法根据计算得到的score值将这个正样本排在负样本前面的概率就是AUC值。所以，AUC值越大，当前的分类算法那越有可能将正样本排在负样本前面，也就是说能够更好的分类。

## 排序的评价方式

第三部分 案例部分

# 分布式推荐系统架构与服务

## 环境准备与软件安装

## 推荐系统的数据流程

预处理

召回

排序

重排序

人工规则

## 实现一个电影推荐系统[[44]](#footnote-44)

### 架构设计

综合业务：Spring、Tomcat

数据存储：业务数据MongoDB、搜索服务数据ES、缓存数据Redis

离线和实施推荐：Spark DF、ML、Streaming

消息服务：Kafka

### 数据加载与处理

数据加载服务，主要用于项目的数据初始化，用于将三个数据集（Movies【电影的数据集】、Rating【用户对于电影的评分】、Tags【用户对于电影的标签】）初始化到Mongodb数据库以及ElasticSearch里面。

【离线推荐】

通过Azkaban周期性的调度【离线统计服务】和【离线推荐服务】。

【离线统计服务】

最热电影统计算法  =>  RateMoreMovies表中

当前最热电影统计算法   =>  RateMoreRecentlyMovies表中

电影的平均评分算法   =>  AvgMovies表中

电影类别TOP10推荐   =>   GenresTopMovies表中。

【离线推荐服务】

通过Spark ALS算法计算模型的训练

基于Model产生电影相似度矩阵   =>  MovieRecs表中

基于Model产生用户推荐矩阵    =>  UserRecs表中

【实时推荐】

当一个用户完成了对电影的评分之后，触发实时推荐，后台服务将评分数据实时写入到LOG日志中。

Kafka会通过kafkaStream程序将log队列中的数据进行格式化，然后将数据推送到recommender的队列中

Spark Streaming实时推荐程序读取kafka中推送过来的数据，配合读取Redis缓存中的数据，运行实时推荐算法【基于模型的推荐】，把结果数据写入到MongoDB的StraemRecs表中

【推荐结果的聚合】

综合业务服务会根据一定的比例聚合 【离线推荐服务（ALS的协同过滤）】、【基于内容的推荐（基于ElasticSearch More like this功能）】、【实时推荐服务（基于模型的推荐）】这些结果。

聚合完成之后，将结果返回到用户端。

前期工作：数据加载

利用Spark SQL将数据分别导入到MongoDB和ElasticSearch

目标

MongoDB

将Movie【电影数据集】/ Rating【用户对电影的评分数据集】/ Tag【用户对电影的标签数据集】数据集分别加载到MongoDB数据库中的Movie / Rating / Tag表中

ElasticSearch

需要将Movie【电影数据集】加载到ElasticSearch名叫Movie的Index中。需要将Tag数据和Movie数据融合。

步骤

先新建一个Maven项目，将依赖添加好。

分析数据集Movie、Rating、Tag

Spark DF加载数据集

将DF加载到MongoDB中：

将原来的Collection全部删除

通过DF的write方法将数据写入

创建数据库索引

关闭MongoDB连接

将DF加载到ElasticSearch中：

将存在的Index删除掉，然后创建新的Index

通过DF的write方法将数据写入

关闭Spark集群

// MONGO\_MOVIE\_COLLECTION，包含电影ID、描述、时长、发行日期、拍摄日期、语言、类型、演员、导演

case class Movie(mid: Int, name: String, descri: String, timelong: String, issue: String, shoot: String, language: String, genres: String, actors: String, directors: String)

// MONGO\_RATING\_COLLECTION，包含用户ID、电影ID、用户对电影的评分、用户评分时间case class MovieRating(uid: Int, mid: Int, score: Double, timestamp: Int)

// MONGO\_TAG\_COLLECTION，包含用户ID、电影ID、标签的具体内容、用户打标签时间case class Tag(uid: Int, mid: Int, tag: String, timestamp: Int)

### 离线推荐

统计推荐

目标

优质电影

获取所有历史数据中评分最多的电影的集合，统计每个电影的评分数 => RateMoreMovies

热门电影

认为按照月来统计，这个月中评分数量最多的电影我们认为是热门电影，统计每个月中的每个电影的评分数量 => RateMoreRecentlyMovie

统计电影的平均评分

将Rating数据集中所有的用户评分数据进行平均，计算每一个电影的平均评分 => AverageMovies

统计每种类别电影的TOP10电影

将每种类别的电影中评分最高的10个电影分别计算出来， => GenresTopMovies

步骤

新建一个项目StaticticsRecommender

统计所有历史数据中电影的评分个数

通过Rating数据集，用mid进行groupby操作，count计算总数

统计以月为单位的电影的评分个数

利用from\_unixtime对timestamp进行转换，结果格式为yyyyMM

通过groupby年月，mid来完成统计

统计每个电影评分得分  
通过Rating数据集，用户mid进行group by操作，avg计算评分

统计每种类别中评分最高的10个电影

需要通过JOIN操作将电影的平均评分数据和Movie数据集进行合并，产生MovieWithScore数据集

利用split和explode将MovieWithScore拆开为以一种Genre为开头的MovieWithScore

此处有两种方案（这里展示第一种，但最好的是建立Aggregator，在ALS离线推荐中介绍）：

通过Genre作为Key，进行groupByKey操作，将相同电影类别的电影进行聚集并按评分取TOP10。（TopN利用PriorityQueue实现）

利用repartitionAndSortWithinPartitions、mapPartition和groupSorted（轮子）实现分Genre的TOP10。

上面数据转化为DF后输出到MongoDB中

离线计算由原来的RDD改为基于Spark DataFrame的实现，而DF本身比RDD具有更多的默认优化，比如Tungsten运用了堆外内存，减少GC开销、序列化比RDD的Java默认序列化更加高效、全代码生成等优势。上述是效率提升是理论上的，下面是实际上的。

首先，DF相比于RDD能这大大地减少了代码量。下面的功能就是简单地统计每部电影的评分数、以月为单位统计每部电影的评分数、每部电影平均分、不同类型电影的评分top10（即分组topn）。这里需要连接电影数据集和评分集，前者包含了电影类型，后者包含平均分。原代码使用RDD的groupByKey，这是一种低效的操作，因为它不会进行map-side聚合，而且分组后使用map，直接调用sortWith对每个分组进行排序，再用slice进行取top10。这种实现并不好，在Scala中运用sortWith虽然方便，但是效率很低，它比toArray.sort慢不少。slice也是，用take更高效。但实际上这里RDD的实现也不需要用全排序再取前10。利用aggregateByKey加优先队列，即堆排的实现，效率更高。其用时由原来的20多秒变为6秒左右。当然，在DF中可用windowfunction来实现topn，但是windowfunction的底层实现是先shuffle然后sort，最后根据所定义的窗口进行相应的聚合。这一过程同样没有map-side聚合，所以尽管数据量少时速度更快，但在数据量大时或许不是一个好的topn实现。其实我觉得最好的topn实现应该是自定义一个aggregator，这个在离线推荐部分会提到。

#### ALS离线推荐

**目标**

训练ALS推荐模型并计算出属于用户各自的Top20电影推荐

计算电影相似度矩阵

**步骤**

训练ALS推荐模型

实例化ALS模型，按常规的模型训练流程训练模型后调用recommendForAllUsers完成推荐。

计算电影相似度矩阵

获取电影的特征矩阵，转换成DoubleMatrix

电影的特征矩阵之间做笛卡尔积，通过余弦相似度计算两个电影的相似度

将数据通过GroupBy处理后，输出

ALS模型的参数选择

通过计算ALS的均方根误差来判断参数的优劣程度

ALS模型需要三列，用户id、电影id和评分，并从中计算出用户矩阵和电影矩阵。项目需要给每个用户推荐20部电影，此处我直接使用了2.2提供的recommendForAllUsers函数便直接得到了结果。但后续项目还要求求出电影的相似度矩阵，用于后续实时推荐计算。这个算法spark没有现成的实现，但是思路根recommendForAllUsers是一样的。就是先求电影矩阵自身的cross join，然后求每部电影对除其自身外的其他所有电影的余弦相似度。公式为两个向量的点积除以两个向量长度的乘积。然后取出每部电影中，余弦相似度前100的电影，这样子构建出电影的相似度矩阵。我根据spark的recommendForAll的源码，重新实现了这一功能，这比原代码的实现快上1倍，而且cross join产生的中间数据的内存也减少了几倍，由几G到了几十M。实际上这里也得出了一个cross join的优化方式.....然后就是取每部电影的前100相似的电影。这里同样是分组TopN，但源码使用了私有的TopNAggregator。同样我也实现了一个，这是我觉得在数据量大时最好的TopN实现。首先它基于DS，也享受了DF的部分优化（比如Catalyst、序列化等）这是优于rdd的aggregateByKey的部分，然后，它包含了map-side聚合，减少了shuffle期间需要传输的数据，这是优于windowfunc的部分。

### 实时推荐

**日志预处理**：构建kafka流，从大量日志数据中过滤出特定前缀的用户评分log，并传到recommender topic让下面的spark streaming接收。

构建LogProcessor实现Processor接口，实现对于数据的处理

构建StreamsConfig配置数据

构建TopologyBuilder来设置数据处理拓扑关系，就是输入、处理器和输出的关系。

构建KafkaStreams来启动整个处理

**前提**

在Redis集群中存储了每一个用户最近对电影的K次评分。实时算法可以快速获取。

离线推荐算法已经将电影相似度矩阵提前计算到了MongoDB中。

Kafka已经获取到了用户实时的评分数据。

读取MonogDB中的相似度矩阵然后广播。读取kafka stream，提取里面的uid、mid、score评分、timestamp，即某个用户对某部电影的进行评分从而产生的日志，然后foreachRDD，对每条日志执行下面三个计算。从被广播的相似度矩阵中获取这次被评分的电影P最相似的K个电影（当然，实现还会从评分集中获取用户已看过的电影，并将这些电影过滤掉）这些作为推荐的候选电影。然后获取用户最近N次电影评分，实际是从redis中读取，这些数据用于调整候选电影的优先级别。后面更多就是利用Scala实现算法的思想了，和spark等技术相关性不大。首先新建一个数组存储候选电影的评分、两个map存储候选电影的增强和减弱因子。然后对每个候选电影进行遍历，遍历中再遍历用户最近评分的电影。针对每个候选电影，看用户最近评分的电影中有没有与它相似的，有的话返回其在相似度表中的分数，没有就返回0.0。然后如果相似度大于0.6，就将这个相似度作为权重，乘以用户对其相似电影的评分。得出候选电影的初步评分。然后根据用户对该相似电影的评分，如果大于3，则候选电影的增强因子+1，否则在减弱因子中-1。最后计算所有部候选电影的初步评分的均值，并加上其增强因子的对数，减去其减弱因子的对数，得出最终得分，并把这个得分表插入到MongoDB用于用户评分后的推荐。

### Spark+ES实时查询

使用 Elasticsearch Spark 连接器将用户事件数据摄入到 Elasticsearch 中并建立索引。

将事件数据加载到 Spark DataFrames 中，并使用 Spark 的机器学习库 (MLlib) 来训练一个协同过滤推荐系统模型。

将训练后的模型导出到 Elasticsearch 中。

使用一个自定义 Elasticsearch 插件，计算个性化的用户和类似电影推荐，并将推荐与搜索和内容过滤相结合[[45]](#footnote-45)。

# 电商平台推荐

## 推荐场景

### 猜你喜欢

猜你喜欢是一种综合性推荐。



### 店铺推荐

当用户点击某个商品时，商品详情页周边和下方显示的本店铺商品推荐。

图片包含 屏幕截图

描述已自动生成

### 组合推荐

商品主页下方推荐的商品配套组合。

图片包含 屏幕截图

描述已自动生成

## 实时推荐lambda架构

大数据的Lambda架构和函数式编程有异曲同工之妙，其核心都是数据的不可变。对不可变的数据进行变形、加工和处理，来回答用户的问题（query）。通过serving layer的预先计算，来降低计算的复杂度。通过speed layer来满足实时性的要求，并且通过把最复杂的设计限制在speed layer来增加系统的可靠性。用函数式编程的观点来设计系统架构，在宏观层面上享受了一些函数式编程的优势。分布式系统需要保证“读”操作的nil-potent和“写”操作的idem-potent。lambda architecture中的元素正好符合该原则query-“读”操作。

只作用于batch和realtime两个view层的数据集，不会修改原始数据。由batch layer来基于原始数据（ground truth）统一掌管“写”操作，发生任何系统故障、意外重启、或发现过去的bug都可以通过一次batch layer升级和重新计算，直接生效。分布式系统中数据一致性的问题是个经典难题，借助batch和speed layer的区分，方便思考、理解该架构下系统一致性方面的特性，通过原始数据、新数据、几种view来明确区分系统中各种数据的相对重要程度，只需要保证原始数据正确安全就好[[46]](#footnote-46)。

### Lambda架构的组件选择



图片包含 屏幕截图

描述已自动生成

Spark Streaming读取Kafka流数据，分为两个不同的工作流进行处理。

第一个部分，将streaming windows size设为5秒，该工作流强调快速响应，不做模型相关的操作。

1. 新用户、新物品的添加

简单的特征计算（抽取keywords，tags）

Redis中各类与此相关的规则数据的计算和更新（比如新品推荐，可能需要把新物品加入到新品推荐列表）

1. 用户行为更细

最新行为保存到redis中

更新用户的特征向量

1. 用户、物品的删除

redis中各类规则数据的计算和更新

第二部分，将streaming windows size设置为1分钟，该工作流可以做模型计算（要求模型的计算在30秒之内完成）

1. 整理数据，尽可能快的更新CF模型
2. 更新CF模型的推荐结果
3. 整理数据，写入增量数据目录，为merge做好数据准别
4. 可以进行比较负责的特征处理

一种简单的实时推荐策略

A．根据用户行为优先计算实时推荐结果

根据用户行为- 推荐品类TopN

B．与离线计算的redis基础推荐列表进行融合

C．优先调用API返回结果Top10（param）

D．存储redis以备用户刷新调取

E．存储mongodb备份(可闲时进行)

2.1实时更新（实时查询）

当用户刷新页面时通过API读取推荐列表实现。比如，一天的推荐结果为Top100，app页面3页显示15条记录。查询更新API读取top15条之后的数据返回给用户。

2.2实时计算

略

## 数据流程

### 数据导入与转换

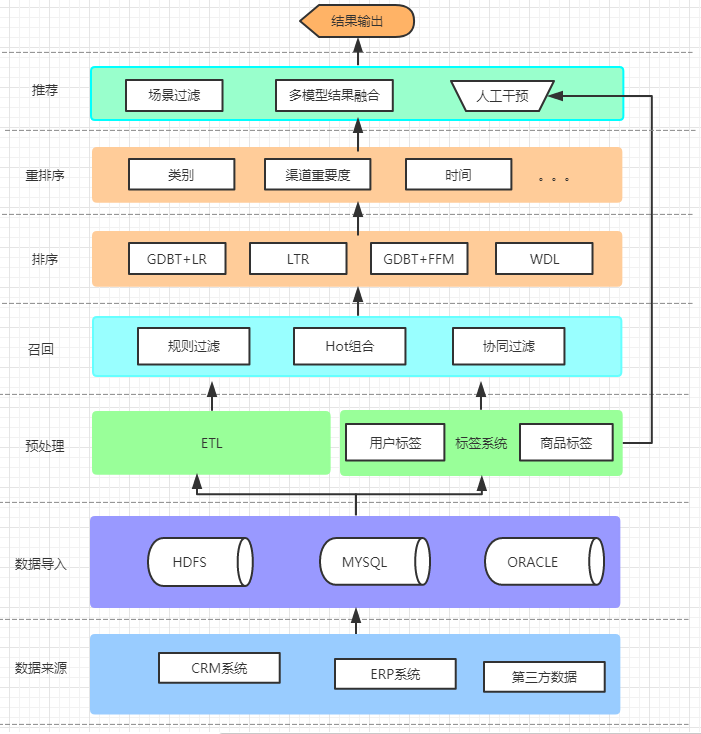
User\_action数据

|  |  |  |
| --- | --- | --- |
| 字段 | 字段说明 | 提取说明 |
| user\_id | 用户标识 |  |
| item\_id | 商品标识 |  |
| behavior\_type | 用户对商品的行为类型 | 包括浏览、收藏、加购物车、购买，对应取值分别是1、2、3、4。 |
| user\_geohash | 用户位置的空间标识，可以为空 | 经纬度 |
| item\_category | 商品分类标识 |  |
| time | 行为时间 | 精确到小时级别 |

商品数据：

|  |  |  |
| --- | --- | --- |
| 字段 | 字段说明 | 提取说明 |
| item\_id | 商品标识 |  |
| item\_ geohash | 商品位置的空间标识，可以为空 | 经纬度 |
| item\_category | 商品分类标识 |  |

上文中很多经典的推荐算法可以通过用户行为或者评分直接获取推荐列表，但在工业界为了提高预测和推荐的准确率、提高响应时间往往会在数据流程上做一些系统性的改良。一般情况下，一个完成批处理流程，主要分四大部分，预处理、召回、排序和重排序。



### 召回（Recall）

召回过程是一种匹配（Match）的过程，也就是基于当前用户（画像、历史行为）和上下文，快速在全库里找到TopN最相关的item，也就是物品的候选集。召回的目的就是减小的排序阶段的物品范围，提高推荐的命中率。召回的方式是比较多的，可以从数据类型还是那个划分，比如基于用户行为和基于用户偏好两种；也可以根据算法划分，比如协同过滤算法和关联算法，不同算法获得候选集再进行融合。

基于用户行为的召回，也就是根据用户购买行为推荐相关或相似的商品，对于已购买的某个商品，不需要重复推荐，可以根据商品种类和用途选择相关或者相似的商品。比如向IPad买家推荐IPad保护套而不是IPad。而对于一些周期性比较强的日用品，像肥皂、洗发水等，则根据一个购买周期再次推荐。

基于用户偏好的召回，也就是用户画像的构建，结合商品品牌、适用人群、价格指数以及用户对商品的点击、购买、关注和收藏等行为。通过用户画像可以确定一些长期推荐的品类。另外，移动互联网背景下，用户习惯于多终端购物，因此需要打通PC、移动app、微信和QQ等信息的融合，从而实现更加精准的画像。

基于地域的召回：一般用户用户冷启动，即网络行为比较少的新用户。

#### 基于Item相似度的召回服务

### 排序（Rank）

排序的步骤一般是对召回的结果进行打分和精细化排序。用户对于与不同商品的接受度是不一样的，排序要做的就是预测商品被用户接受的概率，将概率高的排在展示靠前的位置。排序模块从排序粒度上可以分为精细化排序和粗粒度排序。从指标来看，可以分为单目标排序和多目标排序。因为工业界数据多是隐式反馈类型，不同目标表达了不同的偏好程度，而且用户表达满意度的方式也不尽相同，为了综合目标的收益最大化，一把采用多目标的排序方法。

在融合模型的基础上，推荐排序的问题转化为分类的问题来实现，也就是从用户交互日志中通过模型训练特征权重，再通过LTR排序算法改进，提升转化率。

图片包含 屏幕截图

描述已自动生成

**排序模型**

推荐系统可以理解为，对用户是否会对Item产生某种隐式反馈的分类问题。

推荐列表的获得也就是排序过程，因为我们并不关心单个item预测的评分。

LR => GBDT + LR

图片包含 地图, 文字

描述已自动生成

xgboost内部决策树分裂的部分效果图，f为特征，可以看到树状结构的算法不需要对特征进行归一化，而是根据信息增益大小进行分裂。

通过构建负样本，合并用户交易数据作为训练数据，同时预测用户-商品组合可能购买的概率值。

**LTR**

训练数据：所有用户行为的分组计数 <user,item,count>

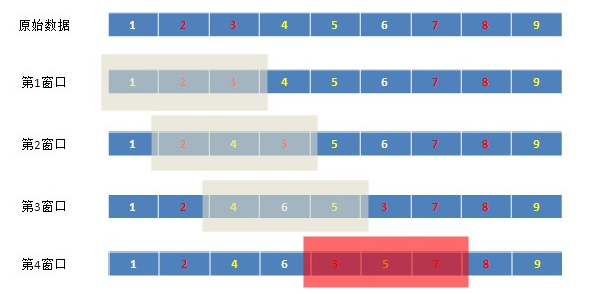
预测数据： 召回集合

xgboost设置objective ： rank:pairwise

### 重排序（Rerank）

重排序可以理解为排序阶段的一种补充性措施，可以有效的防止推荐列表出现同类物品过于密集的现象。比如用户近期一直在看女装外套，如果前十个商品都推荐女装会严重影响客户体验，所以要对推荐物品进行品类的交替和打散，包括品牌的交替和打散。另外，平台在可以在重排序阶段进行强制营销活动（如广告、新品上架）。

下面介绍一种基于滑动窗口的品类打散算法。模拟用户在浏览手机屏幕时的浏览窗口，通过参数设置窗口的大小，保证在窗口向后移动时，窗口内的商品如果出现同一品类，将该商品向后移出该窗口范围。



Scala代码实例：

  /\*\*

• 推荐列表按品类打散

• 品类窗口为w，即相邻W+1个物品不能为同一品类,default:1

• 注意：调用此方法前需要保证品类足够多，否则结果会放弃重复品类的物品

• @return (rank, item)

    \*/

def sortByCate(ori: Array[(String, String)], W:Int=1): Set[(Int, String)] = {

var buffer = new mutable.ListBuffer[(String,String)]

val result = new mutable.ListBuffer[(Int, String)]

for(item <- ori) buffer += item

var windowCate = “”

var index = 0

var i = 1

while (buffer.size > 0) {

if(windowCate == buffer(index).\_2) {

index = index + W

}else{

val temp =  (i, buffer(index).\_1)

result += temp

i += 1

windowCate = buffer(index).\_2 //cate

buffer -= buffer(index)

index = 0

      }

if(index >= buffer.size) buffer.clear()

    }

result.toSet

  }

### 具体Job流程

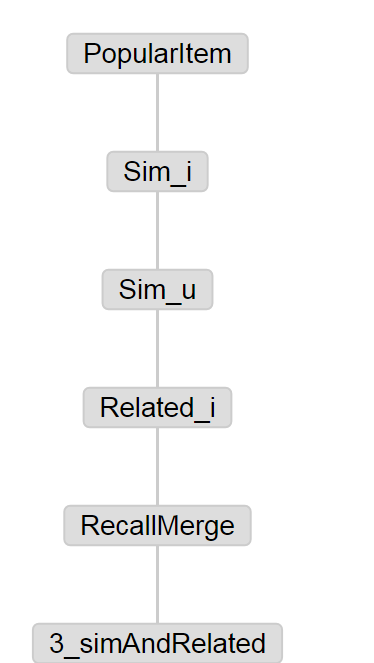
#### 电商数据导入和转换

通过数据格式转换，将电商场景的行为、用户和商品数据合并转换成预设定格式。

图片包含 屏幕截图

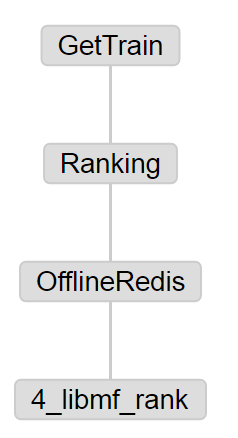
描述已自动生成

#### 召回



#### 精排序

电商场景当前选用了一种简单的方式进行排序。使用libmf进行排序后之间插入redis。



## 各模块实现

[[47]](#footnote-47)根据计算环境的不同，推荐系统的预测大体上可以分为在线（Online）、离线（Offline）两种。在线计算，指的是在线上的推荐服务中，对接受到的请求，进行实时计算，生成推荐结果并直接返回给请求方。离线计算，是指以一定时间周期运行的，对数据库中的大批量数据进行的计算。离线计算的结果通常会写入数据库中，供后续任务读取。除此之外，还有介于在线和离线之间的近线（Nearline）计算，它主要以流处理的方式对近实时的数据进行处理，并将结果写入数据库。

在推荐系统中，在线计算和离线计算有各自的优缺点，及其适合的使用场景。在线计算能够做到实时地对用户行为作出反馈，从而可以针对用户当前所处的环境和临时萌生的兴趣，为其提供更即时、更精准的推荐。但是，受限于系统对于延迟的要求，在线计算必须在算法的复杂性上作出一些牺牲。此外，在线计算能够处理的数据量通常也是比较小的。

离线计算对于算法的复杂性要求则没有那么高。它通常是以批处理的方式在分布式集群上计算。因此，它往往可以处理更大量的数据，考虑更多的特征。但由于离线计算通常是每天一次，因此也就相应的损失了一些实时性，无法对用户行为作出及时反馈。除了预测任务，模型的训练过程也可以算作为一种离线计算。它使用日志系统收集的历史数据，训练得到一个模型，并对其进行性能评估。产出的模型，将会被用于后面的离线和在线预测。模型训练过程，占用的资源多，花费的时间长，比较适合在分布式集群上计算。

离线模型与在线预测之间的融合

一般情况下，离线预测任务可以和模型训练可以完全对接。使用Spark MLlib的Pipeline接口可以将模型训练，以及训练之前的特征预处理、特征工程、特征交叉等阶段，按照一定的顺序组合成一个流水线，同时支持将训练好的整个流水线持久化到磁盘上。在预测阶段，只需要将事先保存的流水线模型加载到集群中，然后将原始特征输入流水线，最终就可以得到模型预测的结果。Spark平台推荐使用DataFrame进行数据训练，分布式环境中进行模型离线预测也可以使用DataFrame。

但是如果放到实时推荐场景中，以上的方法就不管用了，因为它缺乏实时性。实时预测必须要求系统具备实时感知用户实时兴趣，并且做到实时反馈的能力。为了推荐系统实现更好的实时性，需要在线上服务中，使用用户当前的实时特征和反馈，为用户推荐出他当下可能感兴趣的东西。在线上服务中，我们需要对用户可能感兴趣的物品进行排序，使得最合口味的物品被排在推荐列表的前列。因此，需要将离线训练的模型部署到线上服务中。

不同于离线预测，在线的模型预测没有直接使用离线训练存储的流水线模型，而是独立实现了模型预测的算法，以及输入模型前的全部特征预处理过程。因此，每次离线训练结束，将模型部署到线上时，需要将训练得到的模型参数拷贝到线上服务中，同时特征预处理过程需要的参数也需要更新到线上服务中。换句话说，我们需要在线上原封不动地再实现一遍离线训练中定义的特征预处理操作和具体的模型结构。

因此，我们面临了一个两难的困境：既希望模型可以以分布式高吞吐量的方式进行离线训练，同时又希望训练好的模型可以在线上以低延迟的方式进行实时预测。这就是在线预测和离线训练之间的矛盾。

面向实时预测的接口考虑在Spark MLlib的接口上做一些改动，给它加上实时预测的接口。 Spark MLlib的Pipeline接口的通用性，在很大程度上依赖DataFrame这一通用的数据结构。对于Spark MLlib来说，在Pipeline中流动的数据，都是使用DataFrame包装的，每个Transformer都接收一个DataFrame，对其做一个「变形」的操作，然后输出一个新的DataFrame。DataFrame的schema通过Transformer的参数（Param）来约定。

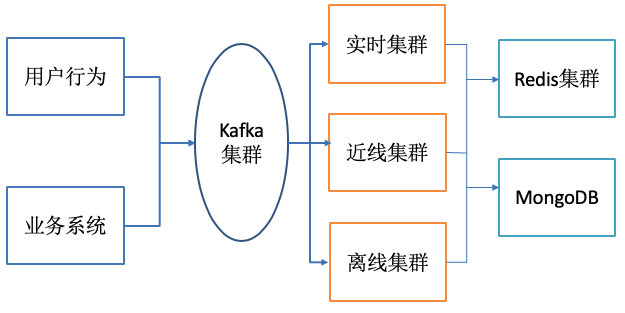
这一套行事方式在分布式计算中非常好用，但在对延迟要求很高的线上服务中，就不太适用了。其中，最主要的原因是，在线上单机计算的环境中将数据转化为DataFrame会有很多不必要的开销，影响服务的延迟。因此，一个直观的想法是：在Spark MLlib中为Transformer提供一个不依赖DataFrame的接口，使其内部核心的计算逻辑直接暴露出来，然后在线上服务中使用这一接口，从而绕过将数据用DataFrame封装这一耗时的操作。

经过观察，我们发现，推荐系统中最常用的模型预测流水线，主要由三个部分组成：特征向量化，特征预处理，和模型预测。其中，特征向量化是将各种数据类型的原始特征转化为向量的过程，它可以由一个特征向量化器（feature vectorizer）完成，其输入是一个某种数据类型的原始特征（比如说一个Map），输出是一个特征向量；特征预处理是对特征向量进行变形、归一化等操作的过程，它可以由一个或多个特征转换器（feature transformer）组成，其中每个转换器的输入和输入都是一个向量类型的特征；模型预测是指将特征输入一个训练好的分类器、回归器，或排序器中，得到一个「分数」的过程，其中「分数」可以表示离散的标签（分类器），也可以表示连续的值（回归器），甚至可以是一个排名（排序器），它的输入是一个向量类型的特征，输出是一个标量值。由于样本特征在存储的时候可能不是采用向量这种类型，而通用的特征处理器都是假设了特征为向量，所以一般我们会在第一步首先将非向量类型的特征转化为向量类型。

基于上面的分析，我们设计了两个面向实时预测的接口，分别用于特征向量化、预处理和模型预测。

预测器的抽象接口是OnlinePredictor。它有一个类型参数FeaturesType，表示这个预测器可以接收的特征类型。我们注意到，对于某些预测器，我们有时需要得到两种类型的预测得分，比如对于二分类器，可能不仅需要输出分类标签，还要输出原始的预测得分。因此，在OnlinePredictor中有predict和predictRaw两个预测接口，根据具体的模型实现需要，可以分别设定两个接口结果的含义。

未完待续。。。



### Kafka模拟数据流

根据用户交互行为产生的日志数据，经过埋点采集，到达kafka进行消息分配。下面模拟以下kafka接收到的埋点数据流。

电商数据的核心字段主要有用户ID、商品ID、用户行为ID、交互时间、商品类别ID。

def simulateECom(spark: SparkSession,  
 producer: KafkaProducer[String, String],  
 params: Params)={  
 val logData = spark.read.textFile(params.source\_input\_path)  
 .*rdd*.map(\_.split(“,”))  
 .map(x => JDataAction(x(0),x(1),x(2),x(3),x(4),x(5),x(6))).  
 filter(x => x.user\_id != “user\_id”)  
 val sample = logData.take(1000000)  
  
 //模拟生成数据  
 for(line <- sample){  
 val event = JSONObject(*Map*(  
 “user\_id” -> line.user\_id,  
 “sku\_id” -> line.sku\_id,  
 “time” -> DateUtil.*getTime*(),  
 “model\_id” -> line.model\_id,  
 “type\_id” -> line.type\_id,  
 “cate” -> line.category\_id  
 // “brand” -> line.brand\_id  
 ))  
 *println*(event.toString())  
 //发送数据  
 producer.send(new ProducerRecord[String, String](params.kafkaInTopic, event.toString))  
 Thread.*sleep*(params.sleep\_time)  
 }  
}

### 数据清洗和特征计算

如上图所示是一个经典的机器学习问题框架图。数据清洗和特征挖掘的工作是在灰色框中框出的部分，即“数据清洗=>特征，标注数据生成=>模型学习=>模型应用”中的前两个步骤[[48]](#footnote-48)。

灰色框中蓝色箭头对应的是离线处理部分。主要工作是

从原始数据，如文本、图像或者应用数据中清洗出特征数据和标注数据。

对清洗出的特征和标注数据进行处理，例如样本采样，样本调权，异常点去除，特征归一化处理，特征变化，特征组合等过程。最终生成的数据主要是供模型训练使用。

灰色框中绿色箭头对应的是在线处理的部分。所做的主要工作和离线处理的类似，主要的区别在于1.不需要清洗标注数据，只需要处理得到特征数据，在线模型使用特征数据预测出样本可能的标签。2.最终生成数据的用处，最终生成的数据主要用于模型的预测，而不是训练。

在离线的处理部分，可以进行较多的实验和迭代，尝试不同的样本采样、样本权重、特征处理方法、特征组合方法等，最终得到一个最优的方法，在离线评估得到好的结果后，最终将确定的方案在线上使用。

另外，由于在线和离线环境不同，存储数据、获取数据的方法存在较大的差异。例如离线数据获取可以将数据存储在Hadoop，批量地进行分析处理等操作，并且容忍一定的失败。而在线服务获取数据需要稳定、延时小等，可以将数据建入索引、存入KV存储系统等。后面在相应的部分会详细地介绍。

本文以点击下单率预测为例，结合实例来介绍如何进行数据清洗和特征处理。首先介绍下点击下单率预测任务，其业务目标是提高团购用户的用户体验，帮助用户更快更好地找到自己想买的单子。这个概念或者说目标看起来比较虚，我们需要将其转换成一个技术目标，便于度量和实现。最终确定的技术目标是点击下单率预估，去预测用户点击或者购买团购单的概率。我们将预测出来点击或者下单率高的单子排在前面，预测的越准确，用户在排序靠前的单子点击、下单的就越多，省去了用户反复翻页的开销，很快就能找到自己想要的单子。离线我们用常用的衡量排序结果的AUC指标，在线的我们通过ABTest来测试算法对下单率、用户转化率等指标的影响。

**特征使用方案**

在确定了目标之后，下一步，我们需要确定使用哪些数据来达到目标。需要事先梳理哪些特征数据可能与用户是否点击下单相关。我们可以借鉴一些业务经验，另外可以采用一些特征选择、特征分析等方法来辅助我们选择。具体的特征选择，特征分析等方法我们后面会详细介绍。从业务经验来判断，可能影响用户是否点击下单的因素有：

距离，很显然这是一个很重要的特征。如果购买一个离用户距离较远的单子，用户去消费这个单子需要付出很多的代价。 当然，也并不是没有买很远单子的用户，但是这个比例会比较小。

用户历史行为，对于老用户，之前可能在美团有过购买、点击等行为。

用户实时兴趣。

单子质量，上面的特征都是比较好衡量的，单子质量可能是更复杂的一个特征。

是否热门，用户评价人数，购买数等等。

在确定好要使用哪些数据之后，我们需要对使用数据的可用性进行评估，包括数据的获取难度，数据的规模，数据的准确率，数据的覆盖率等，

数据获取难度

例如获取用户id不难，但是获取用户年龄和性别较困难，因为用户注册或者购买时，这些并不是必填项。即使填了也不完全准确。这些特征可能是通过额外的预测模型预测的，那就存在着模型精度的问题。

数据覆盖率

数据覆盖率也是一个重要的考量因素，例如距离特征，并不是所有用户的距离我们都能获取到。PC端的就没有距离，还有很多用户禁止使用它们的地理位置信息等。

用户历史行为，只有老用户才会有行为。

用户实时行为，如果用户刚打开app，还没有任何行为，同样面临着一个冷启动的问题。

数据的准确率

单子质量，用户性别等，都会有准确率的问题。

**特征获取方案**

Ok，在选定好要用的特征之后，我们需要考虑一个问题。就是这些数据从哪可以获取？只有获取了这些数据我们才能用上。否则，提一个不可能获取到的特征，获取不到，提了也是白提。下面就介绍下特征获取方案。

离线特征获取方案

离线可以使用海量的数据，借助于分布式文件存储平台，例如HDFS等，使用例如MapReduce，Spark等处理工具来处理海量的数据等。

在线特征获取方案

在线特征比较注重获取数据的延时，由于是在线服务，需要在非常短的时间内获取到相应的数据，对查找性能要求非常高，可以将数据存储在索引、kv存储等。而查找性能与数据的数据量会有矛盾，需要折衷处理，我们使用了特征分层获取方案，如下图所示。

特征工程

出于性能考虑。在粗排阶段，使用更基础的特征，数据直接建入索引。精排阶段，再使用一些个性化特征等。

**特征与标注数据清洗**

在了解特征数据放在哪儿、怎样获取之后。下一步就是考虑如何处理特征和标注数据了。下面3节都是主要讲的特征和标注处理方法

##标注数据清洗

首先介绍下如何清洗特征数据，清洗特征数据方法可以分为离线清洗和在线清洗两种方法。

离线清洗数据

离线清洗优点是方便评估新特征效果，缺点是实时性差，与线上实时环境有一定误差。对于实时特征难以训练得到恰当的权重。

在线清洗数据

在线清洗优点是实时性强，完全记录的线上实际数据，缺点是新特征加入需要一段时间做数据积累。

样本采样与样本过滤

特征数据只有在和标注数据合并之后，才能用来做为模型的训练。下面介绍下如何清洗标注数据。主要是数据采样和样本过滤。

数据采样，例如对于分类问题：选取正例，负例。对于回归问题，需要采集数据。对于采样得到的样本，根据需要，需要设定样本权重。当模型不能使用全部的数据来训练时，需要对数据进行采样，设定一定的采样率。采样的方法包括随机采样，固定比例采样等方法。

除了采样外，经常对样本还需要进行过滤，包括

1.结合业务情况进行数据的过滤，例如去除crawler抓取，spam，作弊等数据。

2.异常点检测，采用异常点检测算法对样本进行分析，常用的异常点检测算法包括

偏差检测，例如聚类，最近邻等。

基于统计的异常点检测算法

例如极差，四分位数间距，均差，标准差等，这种方法适合于挖掘单变量的数值型数据。全距(Range)，又称极差，是用来表示统计资料中的变异量数(measures of variation) ，其最大值与最小值之间的差距；四分位距通常是用来构建箱形图，以及对概率分布的简要图表概述。

基于距离的异常点检测算法，主要通过距离方法来检测异常点，将数据集中与大多数点之间距离大于某个阈值的点视为异常点，主要使用的距离度量方法有绝对距离 ( 曼哈顿距离 ) 、欧氏距离和马氏距离等方法。

基于密度的异常点检测算法，考察当前点周围密度，可以发现局部异常点，例如LOF算法

**特征分类**

在分析完特征和标注的清洗方法之后，下面来具体介绍下特征的处理方法，先对特征进行分类，对于不同的特征应该有不同的处理方法。

根据不同的分类方法，可以将特征分为(1)Low level特征和High level特征。(2)稳定特征与动态特征。(3)二值特征、连续特征、枚举特征。

Low level特征是较低级别的特征，主要是原始特征，不需要或者需要非常少的人工处理和干预，例如文本特征中的词向量特征，图像特征中的像素点，用户id，商品id等。Low level特征一般维度比较高，不能用过于复杂的模型。High level特征是经过较复杂的处理，结合部分业务逻辑或者规则、模型得到的特征，例如人工打分，模型打分等特征，可以用于较复杂的非线性模型。Low level 比较针对性，覆盖面小。长尾样本的预测值主要受high level特征影响。 高频样本的预测值主要受low level特征影响。

稳定特征是变化频率(更新频率)较少的特征，例如评价平均分，团购单价格等，在较长的时间段内都不会发生变化。动态特征是更新变化比较频繁的特征，有些甚至是实时计算得到的特征，例如距离特征，2小时销量等特征。或者叫做实时特征和非实时特征。针对两类特征的不同可以针对性地设计特征存储和更新方式，例如对于稳定特征，可以建入索引，较长时间更新一次，如果做缓存的话，缓存的时间可以较长。对于动态特征，需要实时计算或者准实时地更新数据，如果做缓存的话，缓存过期时间需要设置的较短。

二值特征主要是0/1特征，即特征只取两种值：0或者1，例如用户id特征：目前的id是否是某个特定的id，词向量特征：某个特定的词是否在文章中出现等等。连续值特征是取值为有理数的特征，特征取值个数不定，例如距离特征，特征取值为是0~正无穷。枚举值特征主要是特征有固定个数个可能值，例如今天周几，只有7个可能值：周1，周2，…，周日。在实际的使用中，我们可能对不同类型的特征进行转换，例如将枚举特征或者连续特征处理为二值特征。枚举特征处理为二值特征技巧：将枚举特征映射为多个特征，每个特征对应一个特定枚举值，例如今天周几，可以把它转换成7个二元特征：今天是否是周一，今天是否是周二，…，今天是否是周日。连续值处理为二值特征方法：先将连续值离散化（后面会介绍如何离散化)，再将离散化后的特征切分为N个二元特征，每个特征代表是否在这个区间内。

**特征处理与分析**

在对特征进行分类后，下面介绍下对特征常用的处理方法。包括1.特征归一化，离散化，缺省值处理。2.特征降维方法。3.特征选择方法等。

特征归一化，离散化，缺省值处理

主要用于单个特征的处理。

归一化

不同的特征有不同的取值范围，在有些算法中，例如线性模型或者距离相关的模型像聚类模型、knn模型等，特征的取值范围会对最终的结果产生较大影响，例如二元特征的取值范围为[0，1]，而距离特征取值可能是[0，正无穷)，在实际使用中会对距离进行截断，例如[0，3000000]，但是这两个特征由于取值范围不一致导致了模型可能会更偏向于取值范围较大的特征，为了平衡取值范围不一致的特征，需要对特征进行归一化处理，将特征取值归一化到［0，1］区间。常用的归一化方法包括1.函数归一化，通过映射函数将特征取值映射到［0，1］区间，例如最大最小值归一化方法，是一种线性的映射。还有通过非线性函数的映射，例如log函数等。2.分维度归一化，可以使用最大最小归一化方法，但是最大最小值选取的是所属类别的最大最小值，即使用的是局部最大最小值，不是全局的最大最小值。3.排序归一化，不管原来的特征取值是什么样的，将特征按大小排序，根据特征所对应的序给予一个新的值。

离散化

在上面介绍过连续值的取值空间可能是无穷的，为了便于表示和在模型中处理，需要对连续值特征进行离散化处理。常用的离散化方法包括等值划分和等量划分。等值划分是将特征按照值域进行均分，每一段内的取值等同处理。例如某个特征的取值范围为[0，10]，我们可以将其划分为10段，[0，1)，[1，2)，…，[9，10)。等量划分是根据样本总数进行均分，每段等量个样本划分为1段。例如距离特征，取值范围［0，3000000］，现在需要切分成10段，如果按照等比例划分的话，会发现绝大部分样本都在第1段中。使用等量划分就会避免这种问题，最终可能的切分是[0，100)，[100，300)，[300，500)，..，[10000，3000000]，前面的区间划分比较密，后面的比较稀疏。

缺省值处理

有些特征可能因为无法采样或者没有观测值而缺失，例如距离特征，用户可能禁止获取地理位置或者获取地理位置失败，此时需要对这些特征做特殊的处理，赋予一个缺省值。缺省值如何赋予，也有很多种方法。例如单独表示，众数，平均值等。

特征降维

在介绍特征降维之前，先介绍下特征升维。在机器学习中，有一个VC维理论。根据VC维理论，VC维越高，打散能力越强，可容许的模型复杂度越高。在低维不可分的数据，映射到高维是可分。可以想想，给你一堆物品，人脑是如何对这些物品进行分类，依然是找出这些物品的一些特征，例如：颜色，形状，大小，触感等等，然后根据这些特征对物品做以归类，这其实就是一个先升维，后划分的过程。比如我们人脑识别香蕉。可能首先我们发现香蕉是黄色的。这是在颜色这个维度的一个切分。但是很多东西都是黄色的啊，例如哈密瓜。那么怎么区分香蕉和哈密瓜呢？我们发现香蕉形状是弯曲的。而哈密瓜是圆形的，那么我们就可以用形状来把香蕉和哈密瓜划分开了，即引入一个新维度：形状，来区分。这就是一个从“颜色”一维特征升维到二维特征的例子。

那问题来了，既然升维后模型能力能变强，那么是不是特征维度越高越好呢？为什么要进行特征降维&特征选择？主要是出于如下考虑：1. 特征维数越高，模型越容易过拟合，此时更复杂的模型就不好用。2. 相互独立的特征维数越高，在模型不变的情况下，在测试集上达到相同的效果表现所需要的训练样本的数目就越大。 3. 特征数量增加带来的训练、测试以及存储的开销都会增大。4.在某些模型中，例如基于距离计算的模型KMeans，KNN等模型，在进行距离计算时，维度过高会影响精度和性能。5.可视化分析的需要。在低维的情况下，例如二维，三维，我们可以把数据绘制出来，可视化地看到数据。当维度增高时，就难以绘制出来了。在机器学习中，有一个非常经典的维度灾难的概念。用来描述当空间维度增加时，分析和组织高维空间，因体积指数增加而遇到各种问题场景。例如，100个平均分布的点能把一个单位区间以每个点距离不超过0.01采样；而当维度增加到10后，如果以相邻点距离不超过0.01小方格采样单位超一单位超正方体，则需要10^20 个采样点。

正是由于高维特征有如上描述的各种各样的问题，所以我们需要进行特征降维和特征选择等工作。特征降维常用的算法有PCA，LDA等。特征降维的目标是将高维空间中的数据集映射到低维空间数据，同时尽可能少地丢失信息，或者降维后的数据点尽可能地容易被区分

PCA算法

通过协方差矩阵的特征值分解能够得到数据的主成分，以二维特征为例，两个特征之间可能存在线性关系（例如运动的时速和秒速度），这样就造成了第二维信息是冗余的。PCA的目标是发现这种特征之间的线性关系，并去除。

LDA算法

考虑label，降维后的数据点尽可能地容易被区分

特征选择

特征选择的目标是寻找最优特征子集。特征选择能剔除不相关(irrelevant)或冗余(redundant )的特征，从而达到减少特征个数，提高模型精确度，减少运行时间的目的。另一方面，选取出真正相关的特征简化模型，协助理解数据产生的过程。

特征选择的一般过程如下图所示：

特征工程

主要分为产生过程，评估过程，停止条件和验证过程。

特征选择-产生过程和生成特征子集方法

完全搜索(Complete)

广度优先搜索( Breadth First Search )

广度优先遍历特征子空间。枚举所有组合，穷举搜索，实用性不高。

分支限界搜索( Branch and Bound )

穷举基础上加入分支限界。例如：剪掉某些不可能搜索出比当前最优解更优的分支。

其他，如定向搜索 (Beam Search )，最优优先搜索 ( Best First Search )等

启发式搜索(Heuristic)

序列前向选择( SFS ， Sequential Forward Selection )

从空集开始，每次加入一个选最优。

序列后向选择( SBS ， Sequential Backward Selection )

从全集开始，每次减少一个选最优。

增L去R选择算法 ( LRS ， Plus-L Minus-R Selection )

从空集开始，每次加入L个，减去R个，选最优（L>R)或者从全集开始，每次减去R个，增加L个，选最优(L<R)。

其他如双向搜索( BDS ， Bidirectional Search )，序列浮动选择( Sequential Floating Selection )等

随机搜索(Random)

随机产生序列选择算法(RGSS， Random Generation plus Sequential Selection)

随机产生一个特征子集，然后在该子集上执行SFS与SBS算法。

模拟退火算法( SA， Simulated Annealing )

以一定的概率来接受一个比当前解要差的解，而且这个概率随着时间推移逐渐降低

遗传算法( GA， Genetic Algorithms )

通过交叉、突变等操作繁殖出下一代特征子集，并且评分越高的特征子集被选中参加繁殖的概率越高。

随机算法共同缺点:依赖随机因素，有实验结果难重现。

特征选择－有效性分析

对特征的有效性进行分析，得到各个特征的特征权重，根据是否与模型有关可以分为1.与模型相关特征权重，使用所有的特征数据训练出来模型，看在模型中各个特征的权重，由于需要训练出模型，模型相关的权重与此次学习所用的模型比较相关。不同的模型有不同的模型权重衡量方法。例如线性模型中，特征的权重系数等。2.与模型无关特征权重。主要分析特征与label的相关性，这样的分析是与这次学习所使用的模型无关的。与模型无关特征权重分析方法包括(1)交叉熵，(2)Information Gain，(3)Odds ratio，(4)互信息，(5)KL散度等

**特征监控**

在机器学习任务中，特征非常重要。

个人经验，80%的效果由特征带来。下图是随着特征数的增加，最终模型预测值与实际值的相关系数变化。

特征工程

对于重要的特征进行监控与有效性分析，了解模型所用的特征是否存在问题，当某个特别重要的特征出问题时，需要做好备案，防止灾难性结果。需要建立特征有效性的长效监控机制

我们对关键特征进行了监控，下面特征监控界面的一个截图。通过监控我们发现有一个特征的覆盖率每天都在下降，与特征数据提供方联系之后，发现特征数据提供方的数据源存在着问题，在修复问题之后，该特征恢复正常并且覆盖率有了较大提升。

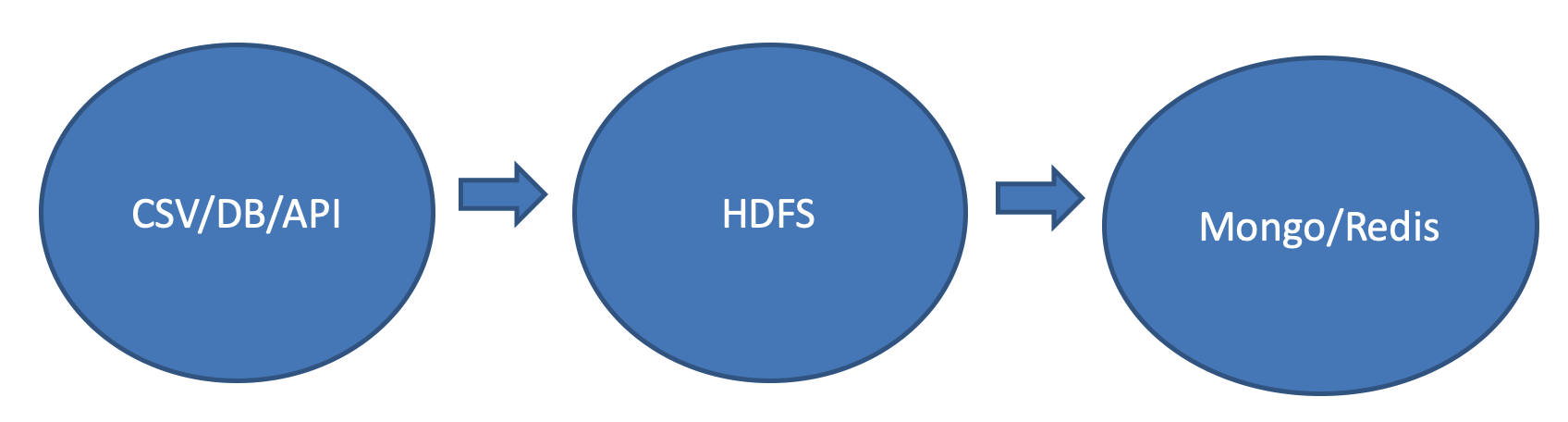
特征工程

在发现特征出现异常时，我们会及时采取措施，对服务进行降级处理，并联系特征数据的提供方尽快修复。对于特征数据生成过程中缺乏监控的情况也会督促做好监控，在源头解决问题。

本文主要介绍了数据清洗与特征处理，其他四篇文章主要介绍了机器学习解决问题流程和模型训练、模型优化等工作。

### 离线计算

离线（offline）模块一般会把几个月的用户数据作为训练数据，更新周期为一天一次。离线计算的数据流支持多种数据源，比如csv文本文件解析、数据库读取或者API调用。一般情况下离线数据会存储半年甚至更久的数据，所以采用HDFS分布式存储。



### PMML模型

Embedding+ 线上简单模型的方法是实用却高效的。但无论如何还是把模型进行了割裂。不完全是 End2End 训练 +End2End 部署这种最“完美”的形式。有没有能够在离线训练完模型之后，直接部署模型的方式呢？本小节介绍一种脱离于平台的通用的模型部署方式 PMML。

PMML 的全称是“预测模型标记语言”（Predictive Model Markup Language, PMML）。是一种通用的以 XML 的形式表示不同模型结构参数的标记语言。在模型上线的过程中，PMML 经常作为中间媒介连接离线训练平台和线上预测平台。

这里以 Spark MLlib 模型的训练和上线过程为例解释 PMML 在整个机器学习模型训练及上线流程中扮演的角色[[49]](#footnote-49)。

spark模型利用PMML的上线过程

例子使用了 JPMML 作为序列化和解析 PMML 文件的 library。JPMML 项目分为 Spark 和 Java Server 两部分。Spark 部分的 library 完成 Spark MLlib 模型的序列化，生成 PMML 文件并保存到线上服务器能够触达的数据库或文件系统中；Java Server 部分则完成 PMML 模型的解析，并生成预估模型，完成和业务逻辑的整合。

由于 JPMML 在 Java Server 部分只进行 inference，不用考虑模型训练、分布式部署等一系列问题，因此 library 比较轻，能够高效的完成预估过程。与 JPMML 相似的开源项目还有 Mleap，同样采用了 PMML 作为模型转换和上线的媒介。

事实上，JPMML 和 MLeap 也具备 sk-learn，TensorFlow 简单模型的转换和上线能力。但针对 TensorFlow 的复杂模型，PMML 语言的表达能力是不够的，因此上线 TensorFlow 模型就需要 TensorFlow 的原生支持——TensorFlow Serving。

Spark离线训练Random Forest模型并保存为pmml格式：

jpmml-evaluator实现在线实时预测：

### 近线计算

近线（nearline）一般一小时更新一次。如果用户操作频发，可以5分钟更新一次。如果用户用户操作行为很少，就可以将近线的行为时间拉长，以获取更大排序候选集。

图片包含 屏幕截图

描述已自动生成

### 实时计算

实时（online）更新，算法要求轻量级，为保证速度，可以通过规则融合的方法。

Spark Streaming实时流处理，将数据流作为一张没有边界的表。

图片包含 屏幕截图

描述已自动生成

# 线下商超案例

线下商超的项目实现了线下商品推荐，对象为华北地区某大型连锁超市。该商场有完整的CRM系统和会员体系，但是痛点在于客户流失率高，很多门店的业绩增加乏力，毛利润不高。通过打通后台CRM会员管理系统，结合客户消费日志挖掘，搭建会员-商品推荐系统。

线下商超的推荐与电商推荐有很大不同，比如交易数据无法做到实时反馈，而且不同电商平台有自己的APP终端，线下商超的网上商城流量很小，主要触达方式是通过微信公众号、小程序推送以及短信提示。

## 需求分析

大型商超的个性化推荐需求总体上分为以下几个步骤：

1. 从客户的基本信息、交易数据中挖掘有价值的信息，并且通过这些信息进一步分析用户的个性化需求。

在推荐系统领域这一步要做的就是充分利用<用户、物品、环境>三要素，计算用户、物品和环境的标签，完善用户和物品画像。其中标签的计算方式多种多样，包括不限于文本、数字、图片等等。

1. 以系统自动计算和分析为主，少量人工为辅，最终提升整体客单价、销售额、毛利润。

## 数据流程

**数据来源**

用户已有的数据接入：CRM，ERP，数据格式可能是各种数据库或者excel文件，csv文件

应支持数据的定时更新。

**数据输入定义**

1. 零售数据 （因为代码已经固定，字段名称不变）

字段不可为空已表明，其他字段可为空字符

|  |  |  |
| --- | --- | --- |
| 字段 | 字段说明 | 提取说明 |
| uid | 订单编号 | not null |
| sldat | 购买时间 | not null |
| vipno | 用户id | not null |
| prodo | 商品id | not null |
| pluname | 商品名称 | not null |
| bndno | 品牌编码 | not null |
| bndname | 品牌名称 | not null |
| qty | 购物数量 | not null |
| amt | 价格 | not null |
| vip\_gender | 用户性别 |  |
| vip\_birthday | 用户出生日期 |  |
| vip\_created\_dat | 会员注册日期 |  |

#### 零售数据导入和转换

线下商超的数据内容和结构与电商基本一致，其中最大的不同在于线下商超场景下顾客没有关于商品的行为类型（action\_type）。取而代之的只有购买行为，也就是说商超日志数据就是顾客的交易记录。

#### ETL与标签系统

原始数据预处理，见step\_0和step\_200-step\_210笔记

etl数据清洗和字段划分

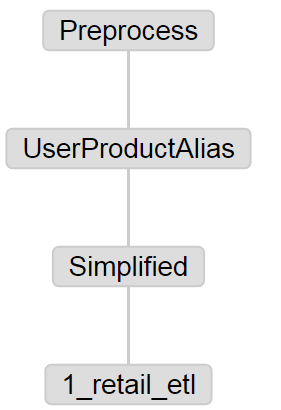
exceptionClean异常清洗

extractAction提取用户行为字段（user\_id, item\_id, rating, time）

extractOperation提取用户操作（user\_id, item\_id, action\_type, category, brand, ...）

extractUserLabel提取用户标签

extractItemLabel提取物品标签



## 推荐场景

### 复购推荐

复购推荐对于线下商超推荐十分重要，因为超市客户的复购比例很高，有些门店复购场景占比超过50%。最常见的复购商品品类包括瓜果蔬菜、柴米油盐、护理产品等等。这些商品有其消耗周期性，比如一个4口家庭可能不到一个月就会买一次酱油，这时推荐系统要做的就是在酱油耗尽前及时提醒客户购买。触达方式上，系统可以发一些满减优惠券，促使客户进行打包消费，从而提高客单价。

复购推荐本身有其特有的复杂度，比如说有些客户定期会购买某一个品牌的酸奶，但是有时会因为商品短缺、注意力转移等等很多因素影响客户的最终成交。

图片包含 文字

描述已自动生成

### 新鲜推荐

给客户推荐之前未购买过的商品。一般场景的推荐都是新鲜推荐，只是商超比较特殊。在超市消费中，周期性消费占比较高，新鲜商品推荐的目的在于提高客单价和整体毛利润。比如对于价格不明感的客户应该适当推荐相似的高毛利商品，但要注意价格的上浮限制。推荐相似商品，就可以运用协同过滤的思想计算商品之间的相似度。

另一种方法与电商场景类似，将新鲜推荐转换为分类或排序问题。先召回候选集，然后根据特征学习的模型预测候选集商品的排名。将排名较高的几个商品通过公众号或者短信进行提示性推荐。

通过客户分群形成的特征空间数据：

图片包含 屏幕截图

描述已自动生成

## 零售业务召回

### 召回策略

召回的具体操作是根据不同的推荐场景决定的。比如给用户推荐重复购买的商品，那召回候选集只能是用户购买超过2次的商品集合。如果是推荐新鲜商品，那召回候选集只能是用户没有购买的商品集合。

对(用户，物品)间关系深入挖掘， 揭示购物模式。

* 1. 复购
  2. 协同推荐
  3. 基于内容的推荐
  4. 大众推荐
  5. 关联推荐

系统把合适的物品推荐给合适的人，促进供需两旺。

* 1. 纯自动化方法
  2. 辅助以人工方法

系统提供合适的界面，展现各种挖掘结果并能够对模型和推荐过程进行一定干预。

**召回方法**

可以根据用户行为、用户画像、显式反馈、隐式反馈

**候选商品**

可以通过，长期行为召回，实时行为召回，热门召回，用户画像召回，搜索词召回，产品词召回。

**过滤规则**

库存过滤，敏感词过滤，人工设定过滤，订单复购过滤

规则过滤

只对交易记录>P的用户做推荐

过滤掉团购用户

过滤掉总交易次数<N的商品，不做为候选集

过滤黑名单（购物袋，季节商品，节日商品）

### 召回模型

#### 人气商品

人气商品实际上是热门商品的匹配，根据不同种类，人气商品可以分为大众热门、群体热门和品类热门等等。人气商品的热度可以进行量化，最简单的就是排名的顺序值，也可以通过特征模型计算。下面实现一个基于简单规则的人气商品排序。

**大众热门**

大众热门就是取所有商品的交易记录，然后给每一个商品计数，最后按照计数大小倒序排序取TopN。大众热门商品一般在新用户加入时使用，因为这个时候系统不知道用户的任何信息。在推荐列表中加入大众热门可以解决部分冷启动问题。

计算大众热门的代码：

//计算大众热门  
val userItemDF = userBehaviorData.  
 map(line => (line.getString(2), line.getString(4))).toDF(“user”,”item”)  
val itemCount = userItemDF.  
 groupBy(“item”).  
 count().  
 orderBy(-*col*(“count”)).  
 map(x=> (transAliasFun(x.getString(0), itemMap), x.getLong(1)))  
//测试  
 itemCount.take(5).foreach(*println*)

一般情况下，线下商超的大众热门靠前的都是蔬菜类，结果输出：

(西红柿,66529)

(后腿肉,60001)

(鸡蛋,59527)

(鲜土豆,45029)

(黄瓜,37581)

**品类热门**

品类下的热门在实际推荐场景用的比价多。当用户的兴趣点在一个品类上，比如“可乐“，根据长尾效应，同一品类下，用户大概率会选择被购买次数较多的那个商品。这也很符合消费心理学，买的多的商品往往在价格和质量上都有很好的保证。

下面是spark计算品类热门的代码：

//计算品类下热门  
case class CateItem(cate: String, item: String)  
def itemCountInList(list: List[String]) = {  
 list.map((\_, 1)).  
 groupBy(\_.\_1).  
 map {case (i, li) =>  
 (i, li.map(\_.\_2).reduce(\_ + *))}.  
 toList.sortBy(-*.\_2)  
}  
  
val itemCate = userBehaviorData.*rdd*.  
 map(line => (line.getString(5), line.getString(4))).  
 groupByKey().  
 mapValues {iter =>  
 itemCountInList(iter.toList).map(\_.swap).sortBy(-\_.\_1)  
 }.flatMapValues(x=>x)  
  
//测试  
itemCate.map{case(user,(count,item))=>  
 (transAliasFun(user,cateMap),  
 (count, transAliasFun(item, itemMap)))}.  
 filter(x=> x.\_1 == “可乐”).  
 take(5).foreach(println)

输出品类“可乐”下面的排名靠前的几个Item：

(可乐,(4328,可口可乐500ml))

(可乐,(4299,可口可乐2.5l))

(可乐,(2032,百事可乐600ml))

(可乐,(2009,可口可乐零度可乐500ml))

(可乐,(1730,可口可乐1.25L+果粒橙1.25L双促装))

**群体热门**

群体热门商品的计算依赖于用户的属性和标签，比如性别、年龄、爱好、习惯等等。多种标签也可以进行组合形成一个群体，比如找出年龄在20-25周岁的喜欢电子产品的男性青年，然后挖掘这类人群的热门商品。

#### 关联商品

**购物篮频繁项统计**

相关商品（购物篮相关，用户相关）

采用频数统计的方式，也就是找出一个购物篮中所有的组合，计算出现频数。

val transactions = userBehaviorData.  
 map(line => (line.getString(0), line.getString(4))).  
 *rdd*.groupByKey().  
 mapValues{items =>  
 val itemPair = items.toArray.distinct.  
 combinations(2).toArray.map(\_.mkString(“|”))  
 itemPair  
 }.flatMapValues(x=>x).map(\_.\_2)

val itemFreq = transactions.  
 map(x => (x, 1)).  
 reduceByKey(\_ + *).  
 sortBy(-*.\_2).  
 map(x=>(x.\_1.split(“\\|”),x.\_2)).  
 map{case(itemPair, count)=> (itemPair(0), itemPair(1), count)}  
itemFreq.map{case(i1,i2,c)=>  
 (transAliasFun(i1,itemMap), transAliasFun(i2,itemMap),c)}.  
 filter(x=>x.\_1.contains(“可乐”)).  
 collect().foreach(println)

上述代码过滤出关于“可乐”的几个频繁项：

输出结果如下:

(可口可乐500ml,翅中,857)

(可口可乐2.5l,翅中,341)

(可口可乐彩罐350ml\*6,雪碧拉罐330ml\*6,292)

(可口可乐1.25L,翅中,253)

(可口可乐零度可乐500ml,翅中,197)

(可口可乐500ml,雪碧500ml,192)

(可口可乐500ml,鲜翅中,190)

根据结果发现频繁项的头部都是交易量最大的蔬菜生鲜类，效果并不好。

**关联规则挖掘**

通过Spark中的关联挖掘类AssociationRule把上述频繁项统计结果作为训练集，置信度阈值取0.3（该值需要根据实际情况进行调节）训练模型。

以下为Spark-Mlib AssociationRules挖掘商超数据关联规则代码：

import org.apache.spark.mllib.fpm.AssociationRules  
import org.apache.spark.mllib.fpm.FPGrowth.FreqItemset  
  
val itemCount = userBehaviorData.  
 map(line => line.getString(4)).*rdd*.map(x => (x, 1)).reduceByKey(\_ + *)  
val transItem = transactions.map(x => (x, 1)).reduceByKey(* + \_)  
val frequency = itemCount.union(transItem)  
val freqItemsets = frequency.map{items =>  
 val itemset = items.\_1.split(“\\|”)  
 val freq = new FreqItemset(itemset, items.\_2)  
 freq  
}

val ar = new AssociationRules().  
 setMinConfidence(0.3)  
val results = ar.run(freqItemsets).cache()  
  
*println*(s”associative itemset count: **$**{results.count()}”)  
val save = results.sortBy(-\_.confidence).map{ rule =>  
 (transAliasFun(rule.antecedent.mkString(“,”),itemMap),  
 transAliasFun(rule.consequent.mkString(“,”),itemMap), rule.confidence)  
}.filter(x => x.\_1.contains(“可乐”))  
save.take(8).foreach(*println*)

打印出“可乐”关键词下的一些关联规则项：

(可口可乐零度可乐330ml,雪碧1.25L,1.0)

(可口可乐零度可乐330ml,芬达2L,1.0)

(可口可乐姜味可乐汽水330ml,农夫东方树叶乌龙480ml,0.6666666666666666)

(百事可乐250ml,好丽友派6P34g\*6,0.3333333333333333)

(百事可乐250ml,芬达橙味250ml,0.3333333333333333)

(百事可乐250ml,梁丰麦丽素巧克力75g,0.3333333333333333)

(百事可乐250ml,南方精装芝麻糊360g,0.3333333333333333)

实际场景中需要对频繁项的结果做一些规则过滤。比如与“可乐”等饮料商品不属于同一个商品大类下的频繁项需要过滤掉。

#### 协同过滤

协同过滤（CF）算法可以利用之前相似度计算的结果进行过滤提取，获得召回结果。CF算法分为基于Item的协同过滤和基于User的协同过滤。一般情况，如果客户的数量远远大于商品的数据量时，建议使用ItemBased CF，一是保证计算效率（商品的维度比用户的维度小），二是保证效果（商品向量相对于用户向量而言没有那么稀疏）。

UserBased CF算法具体的计算方法如下。

根据每个用户的相似度计算结果，对于每个用户U，找到与U最相似度的TopN个用户（N初始化为10）。如果相似度相同，就选择用户ID较小的用户，统计TopN用户所有观看的电影数量，给U推荐K个电影（K初始化取10），如果统计次数相同，选择ID较小的电影。

通过UserBased-CF算法进行召回的代码如下：

def recomToUsers(userSim: RDD[(String, String, Double)],  
 trainData: RDD[(String, String, Double)],  
 recom\_num: Int = 10) = {  
 val similarities = userSim.map{case(u1,u2,s)=>(u1,(u2,s))}  
 // ratings format (user,(item,raing))  
 val ratingsInverse = trainData.map(rating => (rating.\_1, (rating.\_2, rating.\_3)))  
 //statistics format ((user,item),(sim,sim\*rating)),,,,  
 //ratingsInverse.join(similarities) fromating as (user,((item,rating),(user2,similar)))  
 val statistics = ratingsInverse.join(similarities)  
 .map(x => ((x.\_2.\_2.\_1, x.\_2.\_1.\_1), (x.\_2.\_2.\_2, x.\_2.\_1.\_2 \* x.\_2.\_2.\_2)))  
 // predictResult fromat ((user,item),predict)  
 val predictResult = statistics.  
 reduceByKey((x, y) =>  
 ((x.\_1 + y.\_1), (x.\_2 + y.\_2))).map(x => (x.\_1, x.\_2.\_2 / x.\_2.\_1))  
 val filterItem = trainData.map(x => ((x.\_1, x.\_2), Double.*NaN*))  
 val totalScore = predictResult ++ filterItem  
 val finalResult = totalScore.reduceByKey(\_ + \_)  
 .filter(x => !(x.\_2 equals (Double.*NaN*)))  
 .map(x => (x.\_1.\_1, x.\_1.\_2, x.\_2))  
 .groupBy(x => x.\_1)  
 .flatMap(x => (x.\_2.toList.sortWith((x, y) => x.\_3 > y.\_3).take(recom\_num)))  
 *logger*.info(“uBased recom example: “ + finalResult.first)  
 *logger*.info(“uBased recom count: “ + finalResult.count)  
 finalResult  
}

#### 代码目录与输入输出

recall/iSimMatchByRating相似物品计算

i\_sim\_result结果保存磁盘并入库

输入：/xiaoi/recommend/data/etl/train\_implicit

输出：/xiaoi/recommend/data/recall/i\_sim\_result

iSimByRating相似用户计算

u\_sim\_result结果保存磁盘并入库

作为协同过滤提供的候选集

输入：/xiaoi/recommend/data/etl/train\_implicit

输出：/xiaoi/recommend/data/recall/u\_sim\_result

iRelatedMatch相关物品计算

i\_related\_result结果保存磁盘并入库

模块化提供：

1. 相关sku，提供组合推荐

2. 相关品类，如啤酒和尿布

**matchMerge综合所有召回结果**

1. 读取召回文件夹下的所有候选集

2. 去重，过滤黑名单协同过滤目前只获取itembased的结果输入：/xiaoi/recommend/data/recall/

字段： user,item,value

输出：

复购 /xiaoi/recommend/data/recall/reorder

新鲜 /xiaoi/recommend/data/recall/fresh

### 特征计算与特征选择

零售场景目前涉及的字段UID,SLDAT,VIPNO,PRODNO,PLUNAME,DPTNO,DPTNAME,BNDNO,BNDNAME,QTY,AMT,VIP\_GENDER,VIP\_BIRTHDAY,VIP\_CREATED\_DATE

特征计算所需：

UID,SLDAT,VIPNO,PRODNO,DPTNO,QTY,AMT

特征选择

1. 反例的特征没有ui和order特征

2. 基于item的推荐系统，反例组合过于庞大，xgboost单机无法进行，尝试拆分

3. 可以先尝试category的推荐，看看效果是否比LFM推荐的好**新的特征（包含之前的部分特征）** user：

u\_new\_item\_ratio购买新鲜商品所占的比例

u\_new\_category\_ratio购买新鲜品类的比例

u\_new\_m\_category\_ratio购买新鲜中类的比例

u\_new\_b\_category\_ratio购买新鲜大类的比例

u\_time\_new\_item多长时间购买新商品

u\_turnover\_rate用户在验证集中购买的转化率，用户购买的商品个数/用户预测集的商品个数

u\_diff\_hours\_times用户的每两次购买之间的平均时间差，用小时表示

u\_b\_count\_in\_n（n=1/3/6） 用户在观察日前n天的行为总数计数（不同时间粒度：最近1天/3天/6天）

item：

i\_hot\_index商品的热门指数

i\_hot\_in\_small\_category\_index, i\_hot\_in\_middle\_category\_index, i\_hot\_in\_big\_category\_index商品在改小类/中类/大类中的热门指数

i\_small\_category小类

i\_middle\_category中类 （不需要）

i\_big\_category大类 （不需要）

i\_c\_u\_count\_in\_n类别在考察日前n天的用户总数计数 反映了item\_category的热度（用户覆盖性）

i\_c\_b\_count\_in\_n类别在考察日前n天的行为总数计数 反映了item\_category的热度（用户停留性）

i\_u\_count\_in\_n商品在考察日前n天的用户总数计数 反映了item\_id的热度（用户覆盖性） 3

i\_b\_count\_in\_n商品在考察日前n天的行为总数计数 反映了item\_id的热度（用户停留性） 3

i\_bi\_count\_in\_n商品在考察日前n天的各项行为计数 反映了item\_id的热度（用户操作吸引），折射出item\_id产生的购买习惯特点12

i\_b4\_rate商品的点击购买转化率 反映了商品的购买决策操作特点1

i\_b4\_diff\_hours商品的点击购买平均时差 反映了商品的购买决策时间特点1

ui特征： 只有复购模型使用，以及线上电商场景使用

发现：将类别、性别等类别特征转换成one-hot编码，准确率反而降低。可能， Tree-Model不太需要one-hot编码

参考：<https://www.deeplearn.me/1625.html>

## 零售商品排序

排序rank

A复购

B新鲜

LTR排序的结果为分数，而不是概率值。

C热门

D相似商品

E综合模型（多模型融合）

整体式混合

特征组合

特征补充

并行式混合

交叉式混合

其中每个分类采用一种单独的推荐系统。

加权式混合

多个推荐系统结果的加权和。

切换式混合

流水线式混合

串联混合

分级混合

简单电商混合推荐设计：

结果加权（以下所有权重 \* 时间衰减系数）：

1. 主动搜索的权重最高

2. 进入购买订单，未付款情况权重第二。

3. 加购物车（未购买）权重第三。

4. 查看评论，说明有购买意愿

5. 收藏 LR再做一次回归 ：

相关特征的计算，根据用户行为判断其对 热门-复购-新鲜 物品的反应情况。

人工干预（群包匹配）

人工根据用户标签和商品标签批量加入排序列表。

**注意点：**

1. 综合模型是归并多个推荐器的推荐结果，按业务规则进行合并，需要考虑一定的多样性。

2. 每个模型计算得到的<推荐值>的归一化问题。热门模型得到的是权重值，复购模型得到的是概率值，新鲜回归模型得到的是评分值。

3. 目前实现简单的加权融合，考虑两个指标confidence=1-rank/K （K取50或20）， importance = 1 （也就是说默认weight都为1）

零售场景的排序分为复购、和新鲜的子排序，分别计算完成后进行总体的融合操作。最后插入redis。

python/freshRank.py新鲜商品排序

保存磁盘并入库 /xiaoi/recommend/code/src/main/python/fresh\_rank.py

输入：/xiaoi/recommend/data/recall/fresh

输出：/xiaoi/recommend/data/rank/fresh

rank/rankMerge融合模型排序

读取并入库

merge\_model\_result

输入：/xiaoi/recommend/data/rank/reorder

/xiaoi/recommend/data/rank/fresh

输出：/xiaoi/recommend/data/rank/merge

重排序rerank

A品类交替/打散

B品牌交替/打散

C强制营销活动（粒度为排名）

**结果入库database**

各个模型的结果以及最终融合结果都要保存。一般情况下，所有数据都要先放在MongoDB的保存，对于那些请求频繁的用户或者商品可以存放在Redis缓存数据库。

rank/rankInsertDB结果入库

复购和新鲜的推荐结果列表入库

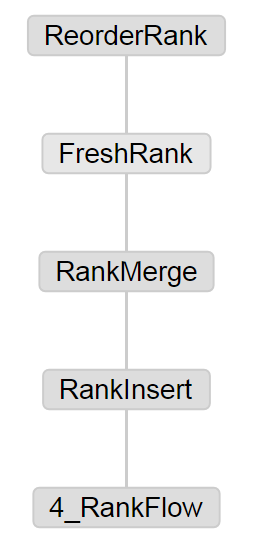
reorder /xiaoi/recommend/data/rank/reorder/prediction\_result.csv

fresh /xiaoi/recommend/data/rank/fresh/

black\_list

white\_list

排序模块总体azkaban Flow流程：



同上召回操作，接受召回结果进入排序模块，分模块排序之后整体融合，最后入库。

# 金融领域的推荐

保险案例为某大型银行保险电销业务。

## 基于客户细分的保险推荐

名单类型分析

图片包含 文字

描述已自动生成

信用卡分群

图片包含 屏幕截图

描述已自动生成

将银行所有的信用卡客户分成四类。

cluster-3：优质客户

cluster-2：潜力优质客户

cluster-1: 一般客户

cluster-4：劣质客户

保险客户分群



构建APP行为的回归模型

图片包含 屏幕截图

描述已自动生成

## Spark GBDT+LR预测高潜客户

基于银行客户的基础信息、账户信息、交易行为、APP浏览行为等等，预测潜在的购买人群。

Adult数据集是UCI机器学习库中数据集，也是机器学习实践的一个很好的入门示例。该数据集记录了若干个人信息和个人年收入。个人年收入与多种因素相关，从直觉上讲，可能受到个人教育程度、年龄、性别、职业等等因素影响。

数据集中一共包含14列属性字段，包含个人的人口统计信息和其他信息。1列类别字段，表示为收入，分为两类：<=50K和>=50K。通过机器学习方法，我们可以根据个人信息预测某个人收入水平的可能性。

数据集下载：<http://www.cs.toronto.edu/~delve/data/adult/desc.html>

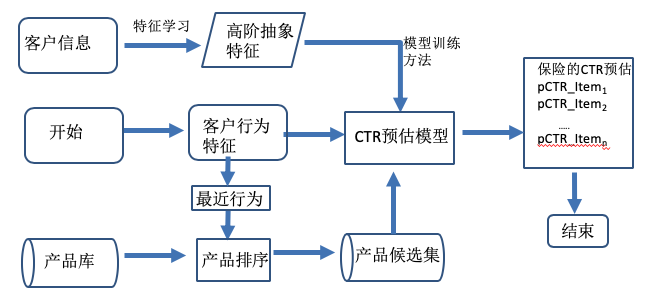
图片包含 文字

描述已自动生成

个人属性包含6个连续属性和8个离散属性。

我们发现数据集中包含7%的缺失值。

## 基于CTR预估的保险需求模型



图片包含 文字, 地图

描述已自动生成

## 集成学习下的融合模型

图片包含 标牌, 天空

描述已自动生成

# 文本推荐

新闻推荐系统研究是当前新媒体时代的热点。在过去，电视、报纸、广播等传统媒体主要是通过人工给受众推荐信息；后来，互联网门户网站和搜索引擎兴起，更多的信息通过热门推荐的方式传递。现在，基于个性化的算法推荐和信息流越来越成为主流。

与电商平台的商品推荐类似，新闻推荐系统是提供个性化的新闻信心，而非商品。所以在算法层面，很多能够在电商平台使用的算法都可以用在新闻推荐中，比如基于用户行为的推荐算法。其中与商品推荐最大的不同在于，新闻推荐系统需要可以采用语义分析和语义理解的方法提供推荐，比如用户在浏览新闻A时推荐与这则新闻主题类似的其他新闻。

## 数据形态

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 唯一编号 | 类别 | 时间 | 浏览次数 | 跟帖次数 | 标题 | 正文 |
| 100000 | 国际要闻 | 2018-12-05 21:05 | 119 | 23 | 国际锐评｜中葡关系翻开新篇章 | 应葡萄牙总统德索萨邀请，中国国家主席习近平星期二（4日）开始对葡萄牙。。。 |

## 热点新闻

新闻的热度值计算，并写入数据库

观看次数\*0.4 + 评论次数 \* 0.5 - 天数时间差\*0.1

import org.apache.spark.sql.{DataFrame, Dataset, Row, SparkSession}  
import org.apache.spark.sql.types.{IntegerType, StringType, StructField, StructType}

import spark.implicits.\_  
  
def readCsv(path: String) = {  
 val newSchema = StructType(*Array*(  
 StructField(“id”, StringType, true),  
 StructField(“cate”, StringType, true),  
 StructField(“new\_time”, StringType, true),  
 StructField(“seenum”, IntegerType, true),  
 StructField(“disnum”, IntegerType, true),  
 StructField(“title”, StringType, true)))  
  
 val data = spark.read.  
 option(“header”, false).  
 schema(newSchema).  
 csv(path)  
 data.as[New]  
}  
  
def calHotValue(data: Dataset[New],  
 seeValue: Float = 0.4f,  
 disValue: Float =0.5f,  
 decayValue: Float = 0.1f) = {  
 val now\_time = DateUtil.*getTime*()  
 data.map(line => {  
 val diff\_days = DateUtil.*dateTimeDiffInDay*(now\_time, line.new\_time)  
 val hot\_value = line.seenum \* seeValue + line.disnum \* disValue - diff\_days \* decayValue  
 (line.id, line.cate, hot\_value)  
 }).toDF(“id”,”cate”,”hot\_value”)  
}

输出新闻ID和热度值:

+------+--------+--------------------+-------------+

| id | new\_id | new\_hot | new\_cate\_id |

+------+--------+--------------------+-------------+

| 1083 | 100000 | 124.2 | 3 |

| 1084 | 100001 | 51.00000000000001 | 3 |

| 1085 | 100002 | 73.10000000000001 | 3 |

| 1086 | 100003 | 164.1 | 3 |

| 1087 | 100004 | 179.5 | 3 |

## 关键词提取

采用TFIDF算法进行关键词提取。TF-IDF是一种用于文本检索和文本转换的常用加权技术，其主要思想就是：如果某个词或者短语在一篇文中中出现的频率TF高，并且在其他文章中很少出现，就认为该词或者短语具有很好的主题区分功能。这个功能同样也适合与文本分类问题上。

假如一篇文件的总词语数是100个，而词语“手机”出现了8次，那么“手机”一词在该文件中的词频就是8/100=0.08。一个计算文件频率 (DF) 的方法是测定有多少份文件出现过“手机”一词，然后除以文件集里包含的文件总数。所以，如果“手机”一词在1,000份文件出现过，而文件总数是10,000,000份的话，其逆向文件频率就是lg(10,000,000 / 1,000)=4。最后的TF-IDF的分数为0.08 \* 4=0.32[[50]](#footnote-50)。

import com.huaban.analysis.jieba.JiebaSegmenter  
import scala.io.Source  
def jieba\_analysis(input: String, path: String): List[String] = {  
 val stopwordPath = path + “stop\_words.txt”  
 val stopWords = Source.*fromFile*(stopwordPath).getLines().toArray  
 val segment = new JiebaSegmenter  
 val seg\_list = segment.sentenceProcess(input)  
 if (stopWords != null) {  
 for(sw <- stopWords){  
 if(seg\_list.contains(sw)) seg\_list.remove(sw)  
 }  
 }else *print*(“Stopwords is null”)  
 seg\_list.asInstanceOf[List[String]]  
}  
  
*/\*\*  
 \* 获取关键词  
 \** ***@param data*** *\** ***@param path*** *\** ***@return*** *\*/*def getKeywords(data: Dataset[New], path: String) = {  
 val splitWords = data.map(file => {  
 (file.id, jieba\_analysis(file.title, path))  
 }).toDF(“id”,”title”)  
  
 import org.apache.spark.ml.feature.{CountVectorizer, CountVectorizerModel, IDF}  
 val cvModel: CountVectorizerModel = new CountVectorizer().  
 setInputCol(“title”).  
 setOutputCol(“rawFeatures”).  
 setVocabSize(1000000). //向量长度  
 setMinDF(2). //词汇出现次数必须大于等于2  
 fit(splitWords)  
 val cvDf = cvModel.transform(splitWords)  
  
 val idf = new IDF().  
 setInputCol(“rawFeatures”).  
 setOutputCol(“features”)  
  
 val idfModel = idf.fit(cvDf)  
 val idfData = idfModel.transform(cvDf)  
  
 val voc= cvModel.vocabulary  
  
 import scala.collection.mutable.ArrayBuffer  
 val keywordsDF = idfData.select(“id”,”features”).*rdd*.  
 map { x => {  
 val name = x.getAs[String](0)  
 //idf结果以稀疏矩阵保存  
 val v = x.getAs[org.apache.spark.ml.linalg.SparseVector](1)  
 var arrW = ArrayBuffer[String]()  
 var arrV = ArrayBuffer[Double]()  
 v.foreachActive((index:Int,value:Double)=>{  
 arrW += voc(index)  
 arrV += value  
 })  
 //根据idf值从大到小排序  
 (name, (arrW zip arrV).toList.sortBy(-\_.\_2).toArray.map(\_.\_1).mkString(“,”))  
 // (name, (arrW zip arrV).toList.sortBy(-\_.\_2))  
 }  
 }  
 keywordsDF  
}

输出新闻ID和对应的关键词：

100000 中葡,葡萄牙,合作,中欧,关系,中国,国家,共建,锐评,首访  
100001葡萄牙,中国,合唱团,葡萄牙人,文化,教学,夫妇,夫妇俩,大学,传播  
100002磋商,经贸,加征,会晤,关税,共识,团队,达成,责编,进晚餐  
100003俄罗斯,刻赤,总统,海峡,俄方,部长,否认,亚速海,秘书,新闻

## 相似新闻

### 基于标签相似的新闻

通过上一步的关键字提取，我们得到了每一篇新闻标题的相关关键词，这些关键词就是新闻的标签。根据这些关键词列表就可以进行两两对比计算，找出标签相似的新闻。

def getSimWithTag(data: RDD[(String, String)], sp: String = “,”) ={  
 val bc\_data = spark.sparkContext.broadcast(data.collect())  
  
 val sim = data.flatMap(rdd\_words => bc\_data.value.map(id\_words => {  
 val bc\_id = id\_words.\_1  
 val bc\_words = id\_words.\_2  
 (bc\_id, rdd\_words.\_1, rdd\_words.\_2, bc\_words)  
 })).filter(x=>x.\_1!=x.\_2).  
 map{case(id1, id2, words1, words2) =>  
 val wordList1 = words1.split(sp)  
 val wordList2 = words2.split(sp)  
 val corr = wordList1.intersect(wordList2).size.toDouble /  
 wordList1.union(wordList2).distinct.size  
 (id1, id2, corr)  
 }  
 sim  
}

输出新闻ID、相似度新闻ID和相似值：

+--------+-------------+------------+-----------------+

| id | new\_id\_base | new\_id\_sim | new\_correlation |

+--------+-------------+------------+-----------------+

| 194677 | 100000 | 100001 | 0.06 |

| 194678 | 100000 | 100022 | 0.18 |

| 194679 | 100000 | 100045 | 0.06 |

| 194680 | 100000 | 100046 | 0.17 |

### 基于用户行为的相似度新闻

基于用户行为的相似度新闻，最简单的就是协同过滤。显然该算法与电商平台和线上商超的商品推荐类似，所以不再赘述。

## 页面展示

主页登录： <http://172.16.20.26:8001/>

提供三个测试用户：

图片包含 屏幕截图

描述已自动生成

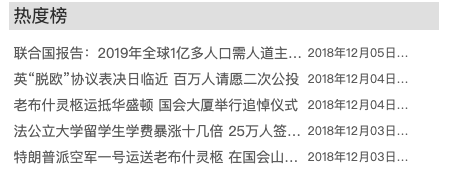
用户冷启动的主题选择（可以不选跳过）：

图片包含 屏幕截图

描述已自动生成

### 推荐场景

**热点新闻**



**主题新闻**

点击不同的主题，按照时间排序，最新的新闻排在前列。

**个性化推荐**

图片包含 屏幕截图

描述已自动生成

个性化推荐计算逻辑：

首先判断用户是否有浏览记录

如果有该用户的浏览记录 则从浏览的新闻获取相似的新闻返回

判断用户最近浏览的新闻是否够10个，如果够的话 取top 10，每个取两个相似

如果不够10个 每个取20/真实个数 +1相似

遍历最近浏览的N篇新闻，每篇新闻取num篇相似新闻

如果该用户没有浏览记录，即该用户是第一次进入系统且没有选择任何标签 返回热度榜单数据的20-40个新闻。

## 新闻推荐的潜在问题[[51]](#footnote-51)

内容不符合用户兴趣

个性化推荐算法并不一定能把符合用户兴趣的内容推荐给用户，造成这个问题的原因有很多。

从数据质量上看，个性化推荐对数据的数量和质量要求比较高，如果一个用户缺乏有质量的数据，个性化推荐远远不如热门推荐的效果。多数新闻资讯类应用要面对大量历史阅读记录空白的新增用户，对他们进行个性化推荐的效果并不好。这个问题被称为“冷启动”问题，是大多数个性化推荐系统面临的共同问题。

从目前自然语言处理的局限来看，个性化推荐系统无法对文章内容做深入理解。目前的自然语言处理只能从其特有高频关键词层面进行标签层面的相似度匹配，这样产生的肤浅话题，无法与用户气质、性格、生活方式进行深层次匹配，很难满足用户阅读新闻资讯中一些细腻的心理需要。

协同过滤算法本身的缺陷，亦导致一些个性化推荐算法的推送内容不符合用户兴趣。一个经典的例子是，娱乐新闻往往很受欢迎，因此用户在协同过滤中的近邻群体多少都阅读过一些娱乐新闻，但这个用户可能从来不读娱乐新闻，强行推荐会使他反感。

内容质量问题

纯粹的个性化推荐系统对文章的质量和内容是没有把关的。算法所做的只是提取文章中包含的特征关键词，并将其与用户兴趣进行匹配。算法并不知道文章水平如何，内容是否健康。一篇文章可能没有任何有意义的内容，只是堆砌一些用户可能感兴趣的词语，但仍然会被算法推送给用户。算法本身是对用户特征和文本数据进行匹配，因此越是相似的结果越会被优先推荐，这容易导致内容同质化。

信息茧房与信息成瘾

信息茧房（Information Cocoons）指的是信息个性化技术使得人们可能减少阅读多样化内容的趋势。由于个性化推荐系统是根据用户已有的阅读偏好进行关键词匹配和推荐，因此相似性较低的内容基本上不会被推荐，这样用户的阅读内容会变得狭隘。在一次又一次阅读自己喜欢领域的信息后，用户不断地在自己与整个世界之间筑起一座高墙。许多人沉湎于这样的拟态环境中，无法自拔。

可遗忘性

个性化推荐算法“记住”了用户看过的文章和相应的喜好，但是对于用户来说，这种记住不一定是好事，用户也存在着让算法忘记他过去喜好的需求。

推荐算法的优化策略

技术不断革新

运用以人工神经网络为代表的新的算法范式，对于推荐系统中许多难以解决的老问题，很可能会有非常好的效果。目前，以深度学习为代表的人工神经网络方法在图像识别、声音识别领域取得了巨大成就，人工神经网络方法，正在被许多研究者尝试运用到推荐系统中。

针对很多新用户缺乏数据沉淀，从而难以进行有效推荐的问题，可以通过获取用户其他平台信息、获取用户已安装的其他软件信息以及引导用户进行口味选择这三种方法来解决。需要注意的是，此举可能侵犯用户隐私，在实际应用中，应当获得用户的许可和授权。

1. <http://xtf615.com/2018/05/03/recommender-system-survey/> [↑](#footnote-ref-1)
2. <https://www.jianshu.com/p/a19486f5a0ea> [↑](#footnote-ref-2)
3. <https://www.cntofu.com/book/85/ml/clean-feature/spark-fextract.md> [↑](#footnote-ref-3)
4. <https://itindex.net/detail/58622-spark-%E7%89%B9%E5%BE%81%E6%8F%90%E5%8F%96-%E9%80%89%E6%8B%A9> [↑](#footnote-ref-4)
5. <http://dblab.xmu.edu.cn/blog/1261-2/> [↑](#footnote-ref-5)
6. <https://blog.csdn.net/qq_34531825/article/details/52415838> [↑](#footnote-ref-6)
7. <https://www.cnblogs.com/massquantity/p/10486904.html> [↑](#footnote-ref-7)
8. <https://blog.csdn.net/u012050154/article/details/60766387> [↑](#footnote-ref-8)
9. <https://zh.wikipedia.org/wiki/Kafka> [↑](#footnote-ref-9)
10. <https://github.com/Snailclimb/JavaGuide/blob/master/docs/database/Redis/Redis.md> [↑](#footnote-ref-10)
11. 按实际情况删减修改 [↑](#footnote-ref-11)
12. <https://www.ruanyifeng.com/blog/2017/08/elasticsearch.html> [↑](#footnote-ref-12)
13. <http://www.zmonster.me/2015/11/15/text_similarity_survey.html> [↑](#footnote-ref-13)
14. 是一种在图形平面上，有多个[节点](https://zh.wikipedia.org/wiki/%E7%AF%80%E9%BB%9E)的[路径](https://zh.wikipedia.org/wiki/%E8%B7%AF%E5%BE%84)，求出最低通过[成本](https://zh.wikipedia.org/wiki/%E6%88%90%E6%9C%AC)的[算法](https://zh.wikipedia.org/wiki/%E7%AE%97%E6%B3%95)。常用于游戏中的NPC的移动计算，或[网络游戏](https://zh.wikipedia.org/wiki/%E7%BD%91%E7%BB%9C%E6%B8%B8%E6%88%8F)的BOT的移动计算上。 [↑](#footnote-ref-14)
15. <https://yongyuan.name/blog/ann-search.html> [↑](#footnote-ref-15)
16. mmapped解释：在计算中，mmap是一个符合POSIX标准的Unix系统调用，它将文件或设备映射到内存中。它是一种内存映射文件I / O的方法。它实现了请求分页，因为文件内容不是直接从磁盘读取的，最初根本不使用物理RAM。在访问特定位置之后，以“懒惰”方式执行来自磁盘的实际读取。在不再需要内存之后，重要的是munmap（2）指向它的指针。可以使用mprotect管理保护信息，并且可以使用madvise强制执行特殊处理。 在Linux，Mac OS X和BSD中，mmap可以创建多种类型的映射。其他操作系统可能仅支持这些操作系统的子集，例如，共享映射在没有全局VFS或I / O高速缓存的操作系统中可能不实用。 [↑](#footnote-ref-16)
17. 一个正版流媒体音乐服务平台，2008年10月在瑞典首都斯德哥尔摩正式上线。 [↑](#footnote-ref-17)
18. <https://github.com/nmslib/hnswlib> [↑](#footnote-ref-18)
19. <https://github.com/microsoft/SPTAG> [↑](#footnote-ref-19)
20. <http://www.zmonster.me/2015/11/15/text_similarity_survey.html> [↑](#footnote-ref-20)
21. <https://github.com/huaban/jieba-analysis> [↑](#footnote-ref-21)
22. <https://blog.csdn.net/u013090676/article/details/81952235> [↑](#footnote-ref-22)
23. <https://blog.csdn.net/u013090676/article/details/82716911> [↑](#footnote-ref-23)
24. <https://spark.apache.org/docs/latest/mllib-feature-extraction.html#word2vec> [↑](#footnote-ref-24)
25. <https://liam-blog.ml/2018/05/04/spark-recommendation-item-based-cf/> [↑](#footnote-ref-25)
26. <https://liam-blog.ml/2018/05/04/spark-recommendation-item-based-cf/> [↑](#footnote-ref-26)
27. <https://www.jianshu.com/p/a3e633e396a0> [↑](#footnote-ref-27)
28. <https://blog.csdn.net/sinat_22594309/article/details/86576757> [↑](#footnote-ref-28)
29. <https://www.cnblogs.com/LeftNotEasy/archive/2011/01/19/svd-and-applications.html> [↑](#footnote-ref-29)
30. <https://www.jianshu.com/p/a3e633e396a0> [↑](#footnote-ref-30)
31. <https://www.cnblogs.com/bjwu/p/9358777.html> [↑](#footnote-ref-31)
32. <https://www.csie.ntu.edu.tw/~cjlin/libmf/> [↑](#footnote-ref-32)
33. <https://blog.csdn.net/canglan211/article/details/89479107> [↑](#footnote-ref-33)
34. 推荐系统原理与实践 213 [↑](#footnote-ref-34)
35. Spark mllib edition2 [↑](#footnote-ref-35)
36. <http://lxw1234.com/archives/2016/01/605.htm> [↑](#footnote-ref-36)
37. <https://www.cnblogs.com/code2one/p/10366183.html> [↑](#footnote-ref-37)
38. <https://zhuanlan.zhihu.com/p/42123341> [↑](#footnote-ref-38)
39. <https://xgboost.readthedocs.io/en/latest/jvm/xgboost4j_spark_tutorial.html> [↑](#footnote-ref-39)
40. <http://xtf615.com/2018/12/25/learning-to-rank/> [↑](#footnote-ref-40)
41. <https://www.cnblogs.com/code2one/p/10343134.html> [↑](#footnote-ref-41)
42. <http://www.voidcn.com/article/p-efnalpyr-sn.html>

    <https://zhuanlan.zhihu.com/p/28067278> [↑](#footnote-ref-42)
43. Spark机器学习第二版 [↑](#footnote-ref-43)
44. <https://www.cnblogs.com/code2one/p/9927026.html#4288560> [↑](#footnote-ref-44)
45. <https://github.com/IBM/elasticsearch-spark-recommender> [↑](#footnote-ref-45)
46. <https://blog.csdn.net/qq_39521554/article/details/79884525> [↑](#footnote-ref-46)
47. <https://zhuanlan.zhihu.com/p/73853438> [↑](#footnote-ref-47)
48. <http://lxw1234.com/archives/2017/11/885.htm> [↑](#footnote-ref-48)
49. [https://www.infoq.cn/article/bLCaNUDP8ztxmlL\*GMwC](https://www.infoq.cn/article/bLCaNUDP8ztxmlL*GMwC) [↑](#footnote-ref-49)
50. <https://www.cnblogs.com/ywl925/p/3275878.html> [↑](#footnote-ref-50)
51. <https://cloud.tencent.com/developer/news/393156> [↑](#footnote-ref-51)