面试题 从 … 羊 蓄水池抽样 Reservoir Sampling

2015年02月24日 11:…

题目：有一个网页抓… …每次调用的时候要等概率的从目前已经抓取的网页中随机选取一… …该怎么实现？

分析：这题题目定义… …先保存当前采集到的所有网页，然后随机采样，这显然不是这题的… …这题想只用O（n）… 其实就等价于有一… …，怎么做随机等概率采样？容易想到的思路是产生一个0到1之间的… …数，然后根据目前… 以随机数构成index… …O（n）存储空间，不符合要求。用bitmap可以把空间减小若干倍… …是还是O（n）的… O（1）空间解决这… …到蓄水池抽样算法，基本道理和从n张彩票中抽奖（假设只有一张… …中奖）无论先抽还… 概率都是1/n类似。 …择（来源这里）

蓄水池算法

对这个问题我们首先… …据。我们接收数据，发现数据流结束了，直接返回该数据，该数据返回的概率为1。看来… 我们试试难一点的情…

我们读到了第一个数… …数据流没有结束。我们继续读取第二个数据，发现数据流结束了。因此我们只要保证以概… 第一个或者第二个数… …个0到1的随机数R,如果R小于0.5我们就返回第一个数据，如果R大于0.5，返回第二个数据…

接着我们继续分析有… …们按顺序给流中的数据命名为1、2、3。我们陆续收到了数据1、2和前面的例子一样，我… 个数据，所以必须… …和上面例子一样，我们按照二分之一的概率淘汰一个，例如我们淘汰了2。继续读取流中… 现数据流结束了，我… …数据3的概率为1/3,那么才有可能保证选择的正确性。也就是说，目前我们手里有1,3两个数… 过一次随机选择，以… …数据1.那么数据1被最终留下的概率是多少呢？

数据1被留下：（1/…

数据2被留下概率：…

数据3被留下概率：…

这个方法可以满足题… …！

因此，我们做一下推… …以1/n的概率留下该数据，否则留下前n-1个数据中的一个。以这种方法选择，所有数据… 选择的概率一样。… …数据被返回的概率都是1/n-1,当前正在读取第n个数据，以1/n的概率返回它。那么前n-… 据被返回的概率为… …这里用到了数学归纳法（中学数学内容），首先证明n=1成立，然后证明当n-1是成立时… 成立，于是对于所有…

所以，简而言之，当… …流到来不断变大）的概率留下当前数，以（n-1）/n的概率留下前面n-1次采样留下的那… 证不断如此采样，当… …样，每个数被采到（留下来）的概率都是1/n.

伪代码如下

```
i = 0
while more inpu…
        with pr…

print choice
```

一个python实现，来自…

```
1  import ran…
2  SAMPLE_CO…
3
4  # Force the…                                    …eatable
5  random.se…
6
7  sample_ti…
8  for index,…                                    …0091103-all-titles-in-ns0")):
9          #…
10         if…
11
12         e…                                     …e)
```

```
13
15
16
17
18
19 print samp
```

```
in the reservoir
14                                  # with a decreasing      ability.
0 and index (inclusive)
dex)

 = line
```

更多蓄水池抽样算法

1 蓄水池抽样—— et/huagong_adu/article/details/7619665

2 Reservoir Samp nts http://gregable.com/2007/10/reservoir-sampling.html

3 几个随机算法 htt n-algorithm.html

最后转载下一篇蓄水 载如下。

数据工程师必知算法

引言：众所周知，想要 工程师——是件很难的事情。我在面试中常使用的方法是：提出即需要算法设计，又需要一些概率论来考察面试者的功底。

“给出一个数据流，这 中数据只能访问一次。请写出一个随机选择算法，使得数据流中所有数据被选中的概率相等。”

当面对这样一个问题的 并没有玩你，相反他可能特别想雇你。他可能正在为无尽的分析请求烦恼，他的ETL流水线已经不在工器学习模型也不再适合 你答出来。

第二件要做的事情是： 的面试官读过Daniel Tunkelang的关于数据工程师的面试建议，那么这个面试题很可能就是他工作中题。所以如果像下面一

“我会首先将输入存到 取结束之后随机选取一个”（大哥，你没看见题目已经说了，数据流长度很大或者未知么，不怕你的下？）

第三件要做的事情是： 体问题（而不是抽象问题），最开始你设计的小例子可能和最后的问题之间相去甚远，但是却能启发解，给你灵感。

蓄水池算法

如前面所说，对这个问 个数据。我们接收数据，发现数据流结束了，直接返回该数据，该数据返回的概率为1。看来很简单，难一点的情况：假设数

我们读到了第一个数据 有结束。我们继续读取第二个数据，发现数据流结束了。因此我们只要保证以相同的概率返回第一个就可以满足题目要求。 我们就返回第一个数据，如果R大于0.5，返回第二个数据。

接着我们继续分析有三 给流中的数据命名为1、2、3。我们陆续收到了数据1、2和前面的例子一样，我们只能保存一个数据1和2中的一个。应该如 之一的概率淘汰一个，例如我们淘汰了2。继续读取流中的数据3，发现数据流结束了，我们知道在长中，如果返回数据3的 就是说，目前我们手里有1,3两个数据，我们通过一次随机选择，以1/3的概率留下数据3，以2/3的概率么数据1被最终留下的

- 数据1被留下
- 数据2被留下
- 数据3被留下

这个方法可以满足题目

因此，我们做一下推论 概率留下该数据，否则留下前n-1个数据中的一个。以这种方法选择，所有数据流中数据被选择的概率证明：假设n-1时候成 正在读取第n个数据，以1/n的概率返回它。那么前n-1个数据中数据被返回的概率为：(1/(n-1))*((n-1)/n)=1/n。设成立。

这就是所谓的蓄水池抽 。你可以在这里找到Greg写的关于蓄水池抽样的算法介绍。本文后面会介绍一下在Cloudera ML中使布式蓄水池抽样和加权

（注：Cloudera ML是

蓄水池抽样在C

分布式蓄水池抽样是G ，基本的蓄水池抽样要求对数据流进行顺序读取。要进行容量为k的分布式蓄水池抽样（前面讨论的我们使用mapreduce 模 合中的每一个元素，都产生一个0-1的随机数，之后选取随机值最大的前k个元素。这种方法在对大数抽样的时候非常管用。 龄，地理信息等的组合。注意如果输入的数据集分布极端的不均匀，那么抽样可能不能覆盖到所有的种分类的组合进行抽样 作纯文本或者hive中的表。

第二个算法更加好玩：
pavlos efraimidis和p
个0-1的值R，并求r的
所以越有可能被返回。

在cloudera ML项目中
化–在输入数据上进行
据一遍就能决定样本组

要读的一些书目

很多有趣的算法都不止
的说，高德纳老爷子的

权重的，算法希望数据被抽样选中的概率和该数据的权重成比例。实际上这之前并不一定有解
ng with a reservoir》。他们的解法既简单又优雅，基本思想和上面的分布池抽样一致：对于每
数据的权重。最终算法返回前k个R值最高的数据然后返回。根据计算规则，大的数据计算所得的

算法），我们会首先使用加权的蓄水池抽样算法对输入数据进行抽样。ksk令会为k-means++
按选入样本的概率和该数据与当前样本中最短距离节点的距离成比例。通过权的蓄水池抽样算法
总代价，之后第二次遍历根据第一次的计算结果进行样本选择）。

，它们有时对于大规模数据分析和一些统计问题也特别有帮助。接下来，我几篇关于算法博客
术》上面的算法吧~

一串巨长的数据流 限制 2024
只需要m+1个空间；假设每个数概率已为m/n;此...

等概率抽样 1882
1、按比例抽样比如按 通过与P比较判断抽样，rando... 来自： 守望者的博客

最全随机抽样算法 1173
1.从N个数中等概率抽 amp;amp;amp;lt;N)是常见的... 来自： bitcarmanlee的博客

等概率随机抽样 1069
蓄水池抽样 问题：要 (N≥k) . 解法：首先从N个... 来自： Owen_1的专栏

海量数据等概率 1040
转自： 问题起源于编 ow could you select one of n ... 来自： WitsMakeMen的专栏

等概率随机取数 749
最近读了项目中的工具 M个数中不重复的随机取出N个... 来自： zuimrs的博客

蓄水池抽样 均匀 616
【转自】http://blog. 如何等概率的从N个元素中选取... 来自： WitsMakeMen的专栏

等概率采样数据流 593
从数据流中等概率的采 ，然后以后的每个数字等概率... 来自： suv1234的博客

随机采样方法 1169
分下面几个部分来介绍 采样 该方法是用一个我们很...

文章热词 信用卡欺 入门到精通 pca降维算法随机参数 数据线性分类

相关热词
bootstrap抽样数据 otstrap抽样方法python c++ 随机两个特定的数据 python面试题培训

关于概率的面试题 3.2万
当前面试中各大名企经 觉得是概率型面试题可以综合... 来自： rudyalwayhere的专栏

关注 排名:227 关注 排名:3000+ 已关 名:167

大数据抽样- 概率
在统计学中，抽样（S ...本（Population，或称为母体）... 来自： CWS_chen 1661

从数据流中随机取
还记得百度网页搜索... ...糊涂，今天恰好找到了这个答... 来自： u012361418的博客 525

随机选择算法
问题描述： 本文主要... ...第k大的数。这个问题最直接的... 来自： DR_eamMer的博客 3330

【算法导论学习-
n个数随机等概率的抽... 来自： BrilliantEagle的专栏 2578

几种等等概率抽样
在统计建模过程中往往会使用到采样技术，通过样本来反映总体特征。关于采样，目前主要有两大类... 来自： wtq1993的博客 2190

解决等概率随机抽
最近在CSDN技术贴区... ...完全随机分给4人，每人可能得... 来自： 感悟编程 913

Apache Flink：
Apache Flink是一个面... ...平台，它能够基于同一个Flink... 来自： 大数据技术杂谈 1.6万

五大开源Web代
Web代理软件转发HT... ...透明代理，而无需客户端配置。... 来自： Enweitech Software ... 6270

负载均衡器部署方
在现阶段企业网中，只... ...设备，较常见是f5、redware... 来自： 淡 7034

Flink架构、原理
Apache Flink是一个面... ...平台，它能够基于同一个Flink... 来自： ppf19159的博客 1164

海量数据随机抽样
随机抽样问题表示如下... ...无法确定。 这种应用的场景一... 来自： Hackbuteer1的专栏 2.5万

亚线性算法-水库
空间亚线性算法：由于... ...放入内存计算，所以一种常用... 来自： zhangvalue的博客 54

数据工程师必知必
摘要： 引言：众所周知... ...——数据工程师——是件很难的... 来自： z69183787的专栏 894

蓄水池采样算法
目录 问题描述分析 算... ...题描述分析 采样问题经常会被遇... 来自： ypfzhao 2650

数据挖掘：大数据
本文转载自：https://... ...67大数据流 即 实时收集的大... 来自： Miss_畅的博客 424

海量数据常见的面
【在海量数据中统计出... ...能够在内存中放下，比如如果... 来自： 碎碎絮語 978

蓄水池抽样问题
问题描述：要求从N个... ...这种应用的场景一般是数据流的... 来自： yinjunshishui的专栏 1183

面试题：随机数生
等概率生成rand5生成... ...杀系统 1639
...的[0, 5)范围内的随机整数，要... 来自： jinzhao1993的博客

蓄水池抽样（Re
在不知道文件总行数的... ...先想到的是我们做过类似的题... 来自： MachileYuan的专栏 339

python实现的四
一、单纯随机抽样（s... ...单位编号，再用抽签法或随机... 来自： 机器学习 2.4万

面试题80：海量
题目： 从N个元素中
个数字被抽中的概率相等。 解读...
1113
来自： tianya_team

数据流基本问题-
在数据流处理中的一个
择一个子集，以便能够对它进...
2386
来自： dm_ustc的博客

随机模拟的基本思
通常，我们会遇到很多
式子特别，真的解不出来； 一...
10.1万
来自： Xianling Mao的专栏

机器学习--蓄水池
学习一下蓄水池抽样以
有限的，随机抽样k个值，可以...
100
来自： 大魔王

蓄水池抽样——
问题：如何随机从n个
但是在此之前你是不知道n的值...
2.6万
来自： Stay hungry, Stay fo...

机器学习-加权采
场景 用户画像的构造
数据，收集用户行为，目的就...
1095
来自： 小码仔

Reservior Samp
蓄水池抽样问题是，从
元素中的每个元素都以相同的概...
917
来自： chouisbo的博客

谷歌面试题：从无
给定一个数据流，其中
索时不断输入的关键字）。如...
（蓄水池抽样）
2836
来自： fanzitao的专栏

flume介绍及原理
flume介绍及扩展开发
ume是Cloudera提供的日志收...
1.4万
来自： wulantian的专栏

Hadoop分布式文
HDFS一流式数据访问
一、HDFS 的概念 1、数据块：...
数据
2262
来自： 俺叫赵小邪的博客

Java字节流与字符
字节流与字符流 先来
的方式进行传输或保存的，程序...
3584
来自： 冷血之心的博客

数字视音频处理知
视频结构化与非线性编
小物理数据单元，所谓镜头检...
2121
来自： (ง •̀_•́)ง

RxJava 并发之惑
Rx 尽量避免状态泄露
状态。比如服务器可以上线和离...
2940
来自： 享受技术带来的快乐

重要性采样（Im
文章转自：http://blo
266 看了好多篇关于采样，重要...
1.7万
来自： 疏影横斜

蒙特卡洛采样之拒
蒙特卡洛（Monte Ca
，在处理一些复杂的问题时成...
2.5万
来自： 白马负金羁

随机抽样算法：蓄
先选取数据流中的前k
amp;amp;amp;lt;= j &a...
166
来自： JackZhangNJU的专栏

matlab随机抽样
1。程序实现 设数据存
及说明如下：m=length(a); %...
1.1万
来自： pretend_b的专栏

有关蓄水池抽样
最近在校论坛上看到了
问题，感觉很有趣，记录如下： ...
401

大数据工程师必备
做大数据的同学经常会
长度很大或者未知。并且对该...
6183
来自： bitcarmanlee的博客

蓄水池抽样算法
蓄水池抽样算法简介 蓄
本中随机选择 K 个样本，其中 ...
1572
来自： 小董的大脑成长园

热门文章

WEKA使用教程(经典教程转载)
阅读量：180774

概率语言模型及其变形系列(1)-PLSA及EM算法
阅读量：68756

概率语言模型及其变形系列(2)-LDA及Gibbs Sampling
阅读量：62173

数据挖掘-决策树ID3分类算法的C++实现
阅读量：53544

Stanford概率图模型（Probabilistic Graphical Model）— 第一讲 贝叶斯网络
阅读量：50998

CarryNLPIR
已关注
原创 粉丝 喜欢 评论
242 2626 50 772
等级： 博客 访问： 149万+
积分： 1万+ 排名： 1400
勋章：

分类
ACM-分治与查找 12篇
ACM-动态规划 41篇
ACM-搜索DFS/BFS 32篇
ACM-数学题 22篇
ACM-Hash 4篇
展开

最新评论
我的微博
计程序判断SVM分类-用JAVA...
weixin_44315211：您好，我的信箱是a22352496@gmail.com，文件可以發我一份嗎，學習一下，謝謝！
Gradient Tree Boo...
weixin_41093521：误人子弟
概率语言模型及其变形系列(5)-L...
qq_40304158：[reply]haungzhuwei[/reply] 我也是遇到同样的问题，请问你们解决了么
概率语言模型及其变形系列(5)-L...
qq_40304158：楼主请问一下迭代次数一定得100么
几句话弄清楚Java参数传值还是传...
olivia77livy：厉害了，的确一句话就说懂了。“当Integer 做为方法参数传递进方法内时，对其的赋值都会导致 原...

联系我们
微信客服 QQ客服
QQ客服 kefu@csdn.net
客服论坛 400-660-0108
工作时间 8:30-22:00

归档
2017年12月 1篇
2017年8月 3篇
2017年3月 1篇
2016年10月 3篇
2016年7月 3篇
展开

android IPC通信 4643
android IPC通信（上 IPC通信（中） - ContentProv... 来自： Shawn_Dut的专栏

vmware9.0安装 fs文件夹 6714
最近在vmware9.0上 与按照vmware tool，但是却无... 来自： tankaro的专栏

在centos 中批量 1502
单机安装cuda ,可以 示一步步的安装。但是当有多... 来自： 草亦花开的专栏

搭建图片服务器 4449
nginx是个好东西，N 服务器，也是一个IMAP/POP... 来自： maoyuanming0806...

Spring Boot M 读写分离 18943
项目地址：https://gi ataSource/tree/dev 在 Sprin... 来自： HelloWood

frp配置本地服务 8290
搭建环境： ubuntu 1 S(阿里云服务器) apache tomc... 来自： Anteoy的博客

.NET和java的RS 26677
RSA .net jva 互通 解 来自： lubiaopan的专栏

ODAC (odp.net 10122
test 来自： 我想我是海 冬天的大...

SSH的端口转发: emote Forward 62453
http://zhumeng833 1172125444948/ 实战 SSH 端... 来自： 明明

jquery/js实现一 16396
jquery/js实现一个网页 多个倒计时. 查阅网络,基本上... 来自： websites

工业相机编程模型 10311
本文详述常见工业相机 要弄懂这里面的过程，基本上... 来自： 文洲的专栏

Hive小文件合并 19220
Hive的后端存储是HD 配置文件系统的块大小，Name... 来自： yycdaizi的专栏

微信支付V3微信 号支付)/JSSDK的使用 3739
扫二维码关注，获取更 对V3微信公众号支付PHP教程/t... 来自： Marswill

linux上安装Doc 89059
最近比较有空，大四出 Docker了，汗汗！ Docker的... 来自： 我走小路的博客

Jmeter:修改内存 13182
在压测过程中jmeter 1、单击压测过程中使用过... 来自： 测试蜗牛，一步一个...

人脸检测工具fac 13477
人脸检测工具face_re 来自： roguesir的博客

plsql的命令（co 20170620 6515
command窗口是命令 lplus命令，基本的命令都可以... 来自： Ape55的博客

寻找连通线,参考 扰线 19804
#include using nam] = {0,1,0,1,0,1,1,0,1,1,0,0,1,0,... 来自： miracle的专栏

3D CNN框架结构 21321
3D CNN框架结构各层 架结构的计算过程，我们都知... 来自： AUTO1993的博客

[转]极线几何约束 7794
来自： volkswageos的专栏

训练网络出现los
4186
粗略写一下： 解决方
态分布，，现在考虑下： http:...
来自： jiachen0212的博客

Two-pass连通域
5106
在Two-pass连通域标
下扫描，会将各个有效像素置...
来自： lichengyu的专栏

加密算法介绍及加
8232
加密算法介绍 一. 密码
置换密码。1881年世界上的第...
来自： leolewin的博客

JAVA结合testng
2680
原理： 1.自已构造一
2.利用testng的监听类在测试...
来自： nicolas_li的专栏

web.config中的
6656
打开某个应用程序
： < sessionState mode="In...
来自： yszwn的专栏

DirectX修复工具
1741116
最后更新：2018-12-
8 增强版 NEW! 版本号：V3.8....
来自： VBcom的专栏

关于SpringBoo
有关）
38665
问题场景描述整个项目
框架一个module spring-boot-...
来自： 开发随笔

内连接，外链接
接大总结
8083
1.什么是连接查询呢？
从这些表中查询数据。 目的...
来自： basycai的博客

强连通分量及缩点tarjan算法解析
21170
强连通分量： 简言之
一个点也是一个连通分量 使用...
来自： 九野的博客

热门文章

WEKA使用教程(经典教程转载)
阅读量：180774

概率语言模型及其变形系列(1)-PLSA及EM算法
阅读量：68756

概率语言模型及其变形系列(2)-LDA及Gibbs Sampling
阅读量：62173

数据挖掘-决策树ID3分类算法的C++实现
阅读量：53544

Stanford概率图模型（Probabilistic Graphical Model）— 第一讲 贝叶斯网络
阅读量：50998

个人分类

ACM-分治与二分查找	12篇
ACM-动态规划	41篇
ACM-搜索DFS/BFS	32篇
ACM-数学题	22篇
ACM-Hash	4篇

展开

最新评论

3行程序搞定SVM分类-用JAVA...
weixin_44315211：您好，我的信箱是a22352496@gmail.com，文件可以發我一份嗎，學習一下，謝謝！

Gradient Tree Boo...
weixin_41093521：误人子弟

概率语言模型及其变形系列(5)-L...
qq_40304158：[reply]haungzhuwei[/reply] 我也是遇到同样的问题，请问你们解决了么

概率语言模型及其变形系列(5)-L...
qq_40304158：楼主请问一下迭代次数一定得100么

几句话弄清楚Java参数传值还是传...
olivia77livy：厉害了，的确一句话就说懂了。“当Integer 做为方法参数传递进方法内时，对其的赋值都会导致 原...

联系我们

微信客服 QQ客服

QQ客服 kefu@csdn.net
客服论坛 400-660-0108
工作时间 8:30-22:00

关于我们 招聘 广告服务 网站地图