阿里电话面试之所做所得所感(2015年7月)

转眼间就到了找工作的阶段，这是我参加的第一个面试，无论结果如何我都受益匪浅。以后也会陆续推出更多的关于面试和找工作相关的文章，希望文章对你有所帮助！而且我准备采用轻松愉快又有内容的方式进行叙述，如果错误或不足之处，还请海涵~

真不敢想象以后成为一名IT男后，每天过着忙碌的上下班挤地铁的生活，晚上回到房间却独守空房，异地他乡的我将如何面对？是啊，很多时候我们的生活都是匆匆忙忙的，都不知道自己到底在做什么！其实正如TED中所言“Stop，Look，Go”,有时候停下来，静静看看思考，感受生活，再继续前行。同样，面试也是一样，我期望每次面试后都能Stop下来，思考总结，再继续准备下一次面试。

一. 面试起因

1.面试起因

说起参加阿里巴巴这次内推过程挺有意思的，起因是我在CSDN写了一篇关于知识图谱的文章：知识图谱相关会议之观后感分享与学习总结，然后有位大哥发私信给我，希望以后多交流并交换了联系方式。后来我们通过QQ成为来了好友，当看到我QQ头像时他惊了个呆(如下图)。在简单交流之后他问我：“在哪里高就？”我说：“今年正准备找工作，研一刚完。”他说：“你想试试淘宝内推吗？”我说：“好啊！”

就这样我参加了我人生的第一次面试，在这里我非常感觉这位大哥，真的谢谢！我一直都对同学说：“我的核心竞争力就是我的人品”，虽然是一句玩笑话，但在我人生中的各个阶段确实都遇到了很多陌生贵人无私的帮助，所以我也经常无私的去帮助陌生人和朋友，无论是生活还是编程上，你也可以试试~

因为我导师的研究方向是数据挖掘和自然语言处理，同时毕业设计在做知识图谱和实体对齐相关的研究，自己对这部分挺感兴趣的，所以申请了“算法工程师”这个职位。

2.职位描述

算法工程师：自然语言处理(NLP)、图像处理、语音识别、机器学习、分布式并行算法、数据挖掘、推荐搜索、复杂网络、深度学习、广告、机器翻译

岗位描述：如何从海量商品中找到最合适的商品、推荐和搜索系统、如何让卖家的商品达到最精确的人群而愿意购买、智能手机、强大云操作系统、商品供应链等

岗位要求：对数据敏感、数学建模、至少会一门编程语言、R语言、机器学习、NLP、图像处理、海量数据处理MapReduce等

简述阿里搜索技术：HBase集群、搜索引擎、资源动态管理、大规模个性化搜索、意图预测、阿里知识图谱（用户/商品多维关系和特色标签用于搜索推荐）、个性化搜索（建立用户Profile、兴趣图谱、关系存储）、文本挖掘技术（分词系统、语义分析、特色标签分析和挖掘）、最大CDN系统、数据库系统、阿里云、淘宝文件系统TFS、Tair、

二. 面试准备

由于参加这个面试很突然，本打算先回家两周陪陪父母，再回来好好准备面试的。而且7月8日我刚考过科目四拿到驾照，7月9日参加毕业开题答辩，7月10日早上10点电话面试，所以准备时间仅仅半天了（6小时），晚上还打了三把dota，哎！我主要准备如下基础东西，而《编程之美》、《剑指offer》也还未阅读，但也希望对你有所帮助吧！

1.数据库

说到面试，经常问的数据库问题就是索引。我准备的问题如下：

题1：数据库中的索引采用什么数据结构？请简述。

索引（index）是一种排序数据结构，为了提高在属性A上查找具有某个特定值的元组的效率，其中Movies(id,name,year,actor)一张电影表的属性就是里面的四个值。它是一棵二叉查找树的键值对，大型关系的索引实现技术是DBMS实现最重要的核心问题。

索引通常使用B树和B+树的数据结构，以协助快速查询、更新数据表中的数据。

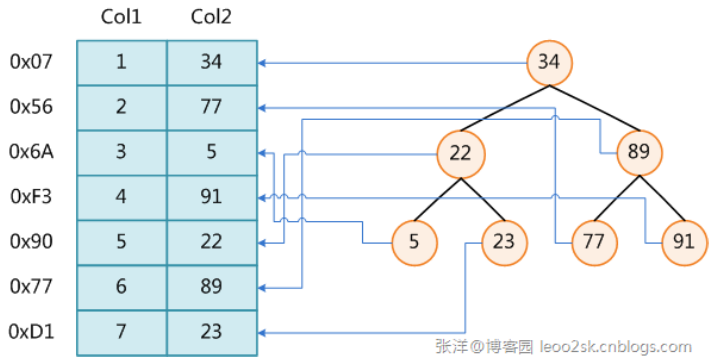
eg: select * from Movies where name='A' and year=1990;

当关系很大时，查询代价太高。若10000个元组需要条件逐个测试，此时可以在Movies和name、year属性上建立索引。

create index keyIndex on Movies(name,year)

详细参考文章：浅谈MySQL索引背后的数据结构及算法-量子恒道

如辉仔的文章"数据库索引的实现原理"所述，下图展示了一种可能的索引方式。



左边是数据表，一共有两列七条记录，最左边的是数据记录的物理地址（注意逻辑上相邻的记录在磁盘上也并不是一定物理相邻的）。为了加快Col2的查找，可以维护一个右边所示的二叉查找树，每个节点分别包含索引键值和一个指向对应数据记录物理地址的指针，这样就可以运用二叉查找在$O(\log_2 n)$的复杂度内获取到相应数据。

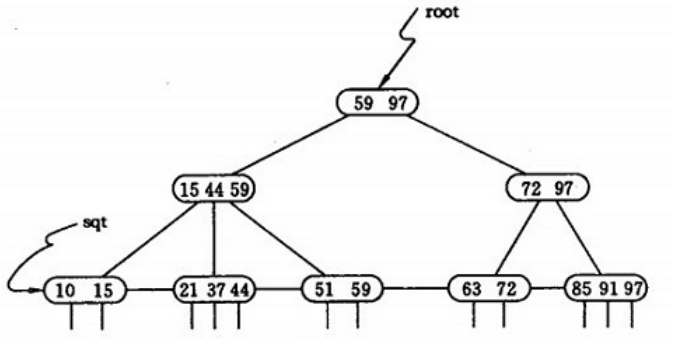
B+树也通常用来做文件索引和文件系统。它的概念如下：参考《数据结构》

(1) 有n棵子树的结点中含有n个关键码

(2) 所有叶子结点中包含了全部关键码的信息，及指向含有这些关键码记录的指针

(3) 叶子结点本身依关键码自小而大的顺序链接

(4) 所有非终端结点可看成索引部分，结点中仅含其子树根结点中最大或最小关键码



如图B+树中有两个头指针，一个指向根节点root，另一个指向关键字最小叶子结点sqt。因此可以对B+树进行两种查找运算：一是从根结点开始随机查找，二是从最小关键字起顺序查找。

重点：索引为什么用B+树就能加快数据检索速度？

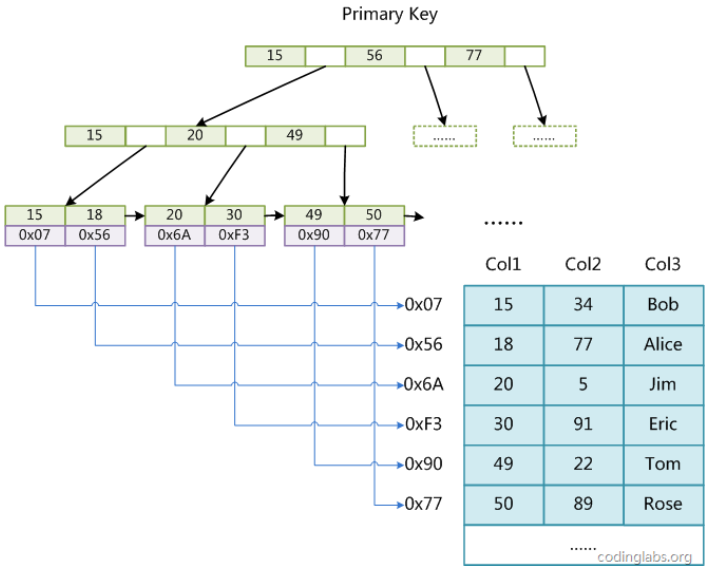
强推：B-树和B+树的应用：数据搜索和数据库索引

索引是对数据库表中一个或多个列的值进行排序的结构。与在表中搜索所有的行相比，索引用指针指向存储在表中指定列的数据值，然后根据指定的次序排列这些指针，有助于更快地获取信息。

通常情况下，只有当经常查询索引列中的数据时，才需要在表上创建索引。索引将占用磁盘空间，并且影响数 据更新的速度。但是在多数情况下，索引所带来的数据检索速度优势大大超过它的不足之处。

二叉查找树进化品种的红黑树等数据结构也可以用来实现索引，但是文件系统及数据库系统普遍采用B-/+Tree作为索引结构。

一般来说，索引本身也很大，不可能全部存储在内存中，因此索引往往以索引文件的形式存储的磁盘上。这样的话，索引查找过程中就要产生磁盘I/O消耗，相对于内存存取，I/O存取的消耗要高几个数量级，所以评价一个数据结构作为索引的优劣最重要的指标就是在查找过程中磁盘I/O操作次数的渐进复杂度。换句话说，索引的结构组织要尽量减少查找过程中磁盘I/O的存取次数。为什么使用B-/+Tree，还跟磁盘存取原理有关。



这里设表一共有三列，假设我们以Col1为主键，图myisam1是一个MyISAM表的主索引（Primary key）示意。可以看出MyISAM的索引文件仅仅保存数据记录的地址。

同时索引的缺点也存在，如创建、维护索引耗时，占物理空间、增删改查维护。在创建索引时应该考虑哪些列（属性）加索引，更方便搜索。

题2：SQL语句(insert\delete\update\select)、事务(rollback\commit)、ACID性质、存储过程(已经执行过，不需再次编译)、触发器(trigger)、BCNF、SQL注入('or'='or')这些基础知识因为当初学得还行，所以就没再看了，也没时间了。

但你应该准备下，尤其是笔试。

2.数据结构

数据结构我也简单过目了下，首先回顾了当初经常问的一个问题"题3"。

题3：数组和链表各自的优缺点?

简述如下：

(1)数组：固定长度，减小内存浪费，方便遍历(通过下标存取)，删除操作后面依次前移，插入操作依次后移，可能遇到超出原定义数组大小，栈分配空间。

(2)链表：动态分配存储，方便增减\插入\删除操作、遍历通过指针依次进行，堆分配空间。

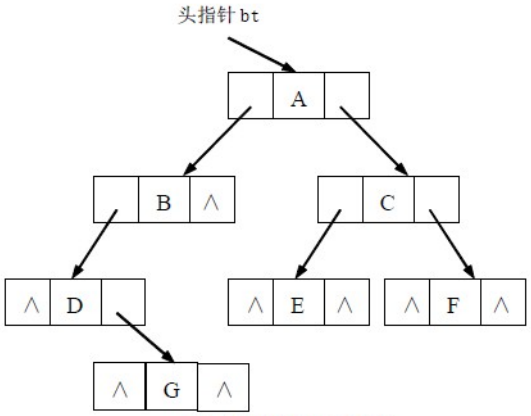
题4：二叉树是个什么鬼？平衡二叉树又是什么?

二叉树主要性质包括：(1)每个结点至多只有两棵子树(不存在度>2的结点)

(2)有左右之分，其次序不能任意颠倒

当时没时间只能回顾概念了，更多二叉树代码强推：轻松搞定面试中的二叉树题目

可采用顺序存储，如 |1|2|3|...|0|8|9| 其中0表示该结点不存在，也可采用链式存储，如下图所示：



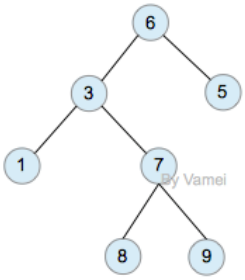
(a) 带头指针的二叉链表

其中有我们非常熟悉的三个遍历——先序遍历(先根、左子树、右子树)、中遍历(左子树、根、右子树)、后续遍历(左子树、右子树、根)。

二叉排序树、平衡二叉树简述之：

(1)左子树!=空，左子树上结点<根

(2)右子树!=空，右子树上结点>根



题5：哈希函数如何解决冲突?

哈希表中的元素是由哈希函数确定的，映射关系。常见哈希函数：

直接地址法： H(key)=key 或 H(key)=a·key+b

除留余数法： H(key)=key MOD p，p<=m

eg: H(key)=key+(-1948) 年份作为关键字加上一个常数。

哈希表：

地址	01	02	03	04
年份	1949	1950	1951	1952
人口	...	...	...	...

建立哈希表需要考虑时间、关键字长度、哈希表大小、分布频率等。

常用的处理冲突方法包括：

(1)开发地址法

开放地址法有一个公式：Hi=(H(key)+di) MOD m i=1,2,...,k(k<=m-1)，其中m为哈希表的长度、di是产生冲突的增量序列。

如果di值可能为1,2,3,...m-1称为线性探测再散列。如果di=1表示每次冲突后向后移动1个位置，还有二次探测再散列和随机探测再散列(di为随机数列)。

eg: 对给定数列{22,41,53,46,30,13,1,67}建立哈希表，表长取9，即[0-8]。哈希函数设定为H(key)=key MOD 8，用线性探测解决冲突Hi=(H(key)+di) MOD m，di=1,2,3,...m-1。(参考百度文库)

解：取22，计算H(22)=22 mod 8 =6，该地址为空，可用：

0	1	2	3	4	5	6	7	8
						22		

取41，计算H(41)=41 mod 8=1，该地址为空，可用：

	0	1	2	3	4	5	6	7	8
		41					22		
比较次数：		1					1		

取53，计算H(53)=53 mod 8=5，该地址为空，可用；

取46，计算H(46)=6，该地址冲突，用线性探测法计算，一个可用地址Hi=(6+1) mod 8=7，该地址为空，可用：

	0	1	2	3	4	5	6	7	8
		41				53	22	46	
比较次数：		1				1	1	2	

取30，计算H(30)=6，该地址冲突，用线性探测Hi=(6+1) MOD 8=7，该地址冲突，再用线性探测计算下一个可用地址，Hi=(6+2) MOD 8=0,该地址为空，可用：

	0	1	2	3	4	5	6	7	8
	30	41				53	22	46	
比较次数：	3	1				1	1	2	

最后建立的哈希表如下所示，其中平均查找长度ASL如下。查找key=67，比较2次找到，查找成功；查找key=21，比较8次找不到，查找失败。

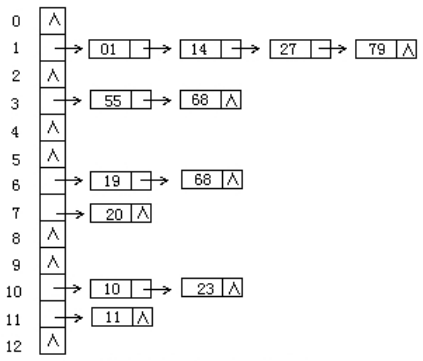
	0	1	2	3	4	5	6	7	8
	30	41	13	1	67	53	22	46	
比较次数	3	1	6	3	2	1	1	2	

平均查找长度ASL= $\frac{1}{8}(3+1+6+3+2+1+1+2)=\frac{19}{8}$

(2)再哈希法

(3)链接地址发

将所有关键字为同义词的记录存储在同一线性链表中。



链地址法处理冲突时的哈希表
（同一链表中关键字有序）

(4)建立一个公共溢出区

题6：图相关知识。最小生成树普里姆算法、最短路径Dijkstra算法、Floyd算法。

PS: 前些天无聊百度了下图灵奖获得者，确实都是些大牛啊！那些XXX算法和XX语言的发明者基本都是其中的成员之一。1978年弗洛伊德图灵奖获得者，Floyd-Warshall算法创始人，但他同时也是堆排序算法、前后断言法的创始人。

3.设计模式

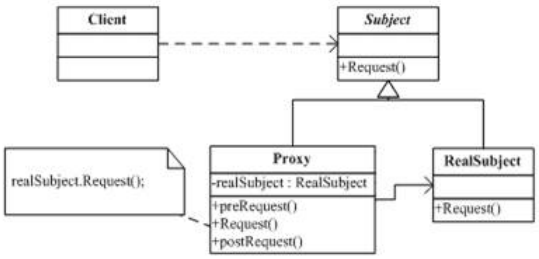
据说可能会问到你熟悉的一个设计模式或项目中使用过的一个设计模式。

题7：简述你熟悉的一种设计模式或项目中用过的设计模式。

代理模式：参考自己的博客"设计模式之代理模式"

其产生原因是有些对象由于某些原因，如对象开销太大、安全保护、远程访问等，直接访问会给使用者或系统带来很多麻烦，所以通过在访问此对象时添加一个对此对象的访问层——代理。如：购买火车票的代售点、银行交易的支付宝等。

代理模式(Proxy Pattern)给某个对象提供一个代理，并由代理对象控制原对象的引用。简言之，一个对象不想或不能直接引用一个对象，可通过”代理“第三者来间接引用，代理对象在客户端和目标对象之间起到中介作用。



举例：一个男孩喜欢上一个女孩，男孩想认识女孩，直接去和女孩打招呼吧，又觉得不好意思，就委托女孩的室友Proxy去帮他搞定这件事，获取女孩的一些信息如QQ、电话、微信。

实际应用：在浏览海量图片时，我看可以使用代理模型，其大图片对应的缩略图就相当于Proxy，当用户喜欢某张具体的图片，点击后在显示具体的大图。否则都显示大图非常浪费资源、内存等。

优点：不让客户直接对对象进行操作，代理可起到保护作用、代理可设置必要的判断、减少系统资源的消耗，对系统进行优化并提高运行速度。

题8：简述设计模式的五大原则。

参考我的文章"设计模式之SOLID原则再回首"，主要包括：

(1) 单一职责原则(Modem，一个类只干一件事拨号\挂机、发送\接受请求)

(2) 开闭原则(增加新功能，原有功能代码关闭，如计算器运算类，子类加法类、减法类)

(3) 里氏替换原则(企鹅继承鸟，但企鹅不会飞，使用父类也适用于子类)

(4) 接口隔离原则(多个和客户相关的接口要好于一个通用接口)

(5) 依赖倒置原则(高层次模块不应该依赖于低层次的模块，二者都应该依赖于抽象)

设计模式的核心思想是通过增加抽象层，把变化部分从不变化部分分离出来。

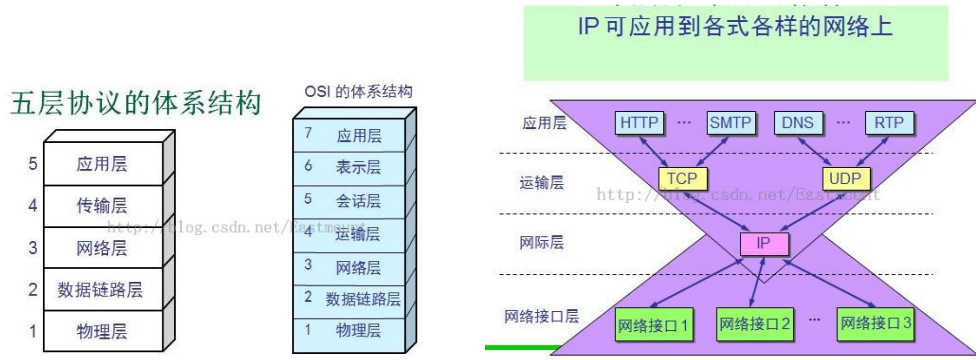
题9：请简述工厂模式和抽象工厂模式优缺点。

这个就没有深入看了，它们是很常用的两种设计模式。

4.计算机网络

题10：请简述计算机网络的五层协议。

下图非常重要，就围绕它们讲即可。



题11：请简述TCP/IP三次握手。

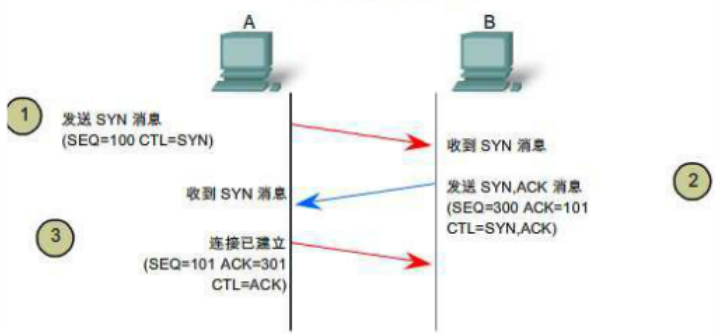
如下图所示，简单打个比方：

(1)A发请求数据包:"我想发数据给你，可以吗?"第一次对话

(2)B发送同意连接，要求同步"可以，你什么时候发?"第二次对话

(3)A再发一个确认主机B的要求"我现在就发，你接着吧"第三次对话

经过三次对话后，主机A向主机B发送数据。



题12：请简述Socket套接字C/S建立过程。

因为我自己做过很多C#相关套接字的客户端/服务器通信的程序(你可能不会遇到这种问题)，所以自己也准备了这个相关知识点。Socket套接字包括IP地址、主机、Tcp协议，bind方法绑定端口、listen方法监听端口、accept接收请求、connect请求链接、send方法发送、receive方法接收、close关闭连接、shutdown停止监听。这就是整个通信的简单过程。

详见：C# 网络编程之套接字编程基础知识

题13：TCP和UDP的区别。

(1)TCP：传输控制协议、基于连接、可靠连接、安全

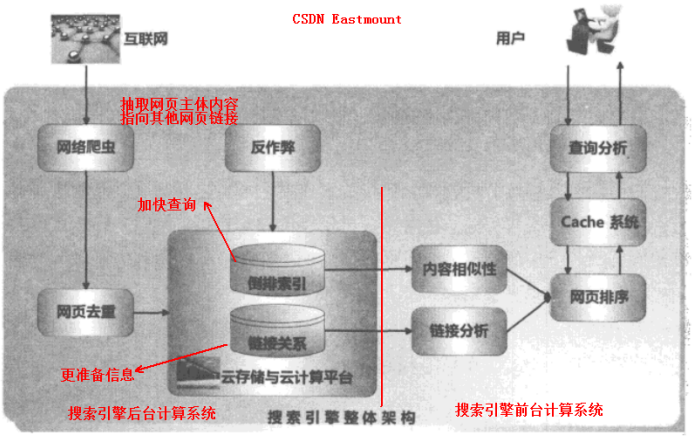
(2)UDP：用户数据报协议、面向非连接、不与对方建立连接、速度快

5.搜索引擎和推荐系统

因为我面试的是搜索部门，所以还是需要知道搜索相关的一些知识。同时很多面试还是会简单问问搜素相关的内容，可能对你有用。

题14：请简述搜索引擎如何实现的?

通过《这就是搜索引擎——张俊林》这本书的一张图即可简单叙述。



搜索引擎后台：

从互联网中抽取网页内容及指向其他页面的链接，因为互联网中有相当比例的内容是完全相同或近似重复的，网页去重模块就是做出检测并去掉重复内容。之后网页内容通过倒排索引这种高效查询数据结构来存储，而网页链接关系也会保存，因为链接分析可帮助用户获取更准确的搜索结果。

由于网页数量太多，同时需要保存一些中间的处理结果，Google提供的云存储与云计算，使用数以万计的普通PC作为海量信息搜索引擎，它成为了搜索引擎的基础支撑。

搜索引擎前台：

当搜索引擎接受到用户的查询词后，首先对查询词进行分析，希望能够能够结合查询词和用户信息来正确推导用户的真正搜索意图。

先在缓存中查找，搜索引擎的缓存中存储了不同的查询意图对应的搜索结果，如果存在则直接返回，即节省资源又加快响应速度；如果没有找到则调用网页排序模块，根据用户查询实时计算哪些网页满足用户信息需求，并排序输出作为搜索结果。两个参考因素：内容相似性和链接分析。

PageRank链接分析提高了搜索质量，它帝归每个网页节点的得分，直到稳定。

题15：推荐系统相关的知识。

(PS：这部分研究较少，只能纸上谈兵，后面面试中会讲到)

分为三大模块：用户建模模块、推荐对象模块和推荐算法模块。

用户建模模块根据用户偏好，TF-IDF算法。

推荐对象模块根据推荐对象特征提取，每个推荐特征的影响、是否自动更新等，SVM分类算法，例如音乐推荐。

推荐算法模块是最核心、关键的模块。(1)基于内容的推荐方法，如音乐共性、兴趣点，根据推荐对象内容特征和用户模型兴趣特征计算相似性，最简单的就是余弦距离计算方法。(2)协同过滤推荐，如邮件系统、书籍，自己身边朋友都选择购买，自己很大概率也会购买。

个性化推荐系统，根据用户的信息需求、兴趣将感兴趣的产品推荐给用户。海量推荐如何更准?

题16：关于淘宝搜索相关知识。

此次强烈推荐阅读《你在淘宝上买了一件东西》这篇文章，也可以通过我的博客阅读这个故事。链接如下：

《淘宝技术这十年》读书笔记 (一).淘宝网技术简介及来源

简单叙述过程如下：

DNS服务器，先把taobao.com转成IP

不同网络转换不一样，就会涉及到负载均衡找到一个更快的IP入口

此时产生一个PV页面访问量

访问生成页面分配给其中一台Server，涉及到公平公正平均，LVS负载均衡完成

逻辑运算和数据处理后，淘宝网首页HTML内容生成

浏览器加载CSS、JS、图片、脚本、文件资源

资源分布多个域名，浏览器同一个域名并发加载数量有限

CDN内容分发网络分配到最近结点(全国各地)，保证淘宝访问海量数据的速度仍然未减慢

CDN同步共用，TFS淘宝分布式文件系统

加载完首页后，搜索框中输入：毛衣，产生一个PV

分词处理(中文"学生"字单位，英文"student"词单位)

搜索购物意图分析
交易记录、搜索记录日志记录，作为后续数据分析
根据用户的意图推荐产品

6.海量数据处理

随着分布式、并发处理、云计算、大数据、MapReduce、Spark等新兴，海量数据问题也是常常被问到的内容，我也简单做了准备。

大家都知道我会强推July的文章：教你如何迅速秒杀掉：99%的海量数据处理面试题

题18：(腾讯)在40亿个海量数据中如何判断一个是否存在?

申请512M内存，一个bit位代表一个unsigned int值，读40亿个数，设置相应bit位，读入要查询位，查看bit是否为1，若为1表示存在否则表示不存在。

题19：海量数据处理。

简述July的处理方法。

时间：Bloom filter、Hash、bit-map、堆、倒排索引

空间：分而治之、hash映射

集群：分布式、并行计算

例：300万个查询字符串中统计最热门的10个查询

若10亿个则先划分小，1000个小文件中，再hashmap通过数量，归并top10，而该题数据300万较小，古内存中处理，HashTable+堆实现<key,value>对应<字符串,次数>的top10获取。

7.NLP和LTR

最后说说数据挖掘、机器学习和NLP（Natural Language Processing，自然语言处理）、LTR(Learning To Rank，学习排序)相关的知识。

题20：请简述机器学习、NLP、LTR相关知识。(自己相关)

传统是按照指令一步一步执行，而机器学习不接受输入的指令，只接受输入的数据。例如：小Y约会经常迟到，预测这次会不会迟到。它会根据已有的数据(经验)训练出模型，在输入新的数据进行预测未知属性。再如NG教授说的根据线性回归预测某个点房屋大小的房价、预测肿瘤是否良性等。

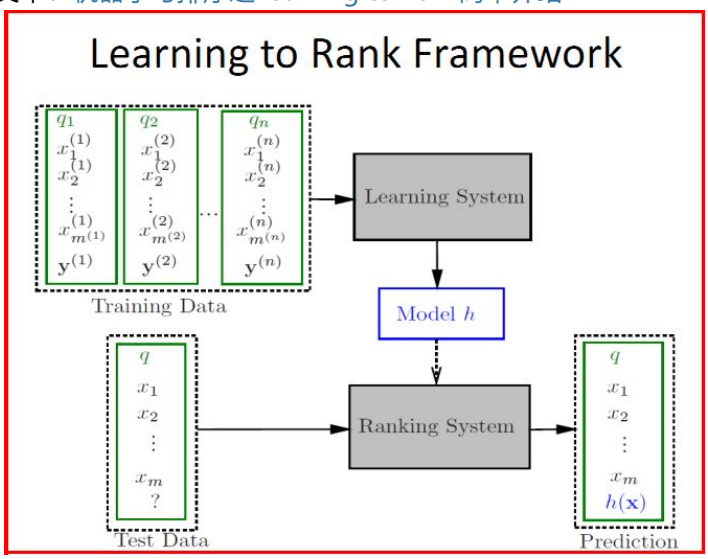
推荐资料：机器学习科普文章：“一文读懂机器学习，大数据/自然语言处理/算法全有了”

数据挖掘=机器学习+数据库

NLP=机器学习+文本处理

计算机视觉=机器学习+图像处理，例如手写识别字

LTR如下图所示，推荐文章：机器学习排序之Learning to Rank简单介绍



三. 面试过程

PS：由于准备时间只有6个小时，中间还打了几把dota2，所以准备的东西就只有这些了。然而这些东西回过头来看并没有什么用，其实就相当于和面试官进行了一场”裸聊“吧！但对我的收获还是非常大，非常感谢他和推荐人，谢谢！OK，剧情开始。

时间：2015年7月10日早上10点半

地点：某某大学计算机实验室四层楼梯

场景：屋外暑假装修”轰隆隆“的施工声

准备：我早上9点起的床，担心错过阿里的电话（确实阿里昨天约好的10点半，非常准时），来到了实验室，同时简单的看了些以前写过的博客，因为在简历中我简单提到了我个人喜欢分享，写了2年多的博客，所以简单过了几篇。快到10点半的时候，去到了楼梯处，手里拿着我自己的一份简历，因为他肯定会问你简历相关的内容。同时由于暑假学校很多装修施工，屋外非常吵，我也没有换个地方，距离太远。

强烈建议你自己面试时找个安静的地方，一个是自己不被影响，另一个是对面试官的尊重吧！他也能听清楚。

过程：

10点30分 电话响起，是杭州来的电话。

面试官：您好！我是淘宝搜索部门的面试官，请问您是XXX，对吧！

作者：您好！我是学生XXX。

面试官：我看了你的简历，你是做过LTR相关的项目，对吧！

点评：所以说前面我准备的很多都没有什么用，面试官上来就会直接问你所投部门和相关专业的知识，而不是基础知识，因为他可能默认你已经会了吧，这第一次面试也让我认识到了这点。但是如果你想参加笔试或者面试实习，你非常有可能会遇到上面我所准备的基础问题，再或者有可能你的面试就会遇到，看看也并没什么错！

作者：对，我们课程做过这方面的大作业。

面试官：那你能简单说说LTR你们是怎么做的吗？

作者：Learning To Rank学习排序，随着海量数据规模越来越大，传统的搜索引擎是通过用户输入的关键字，获取相关内容和链接分析作一个结果排序，返回给用户。而随着特征越来越多，好像Google现在就200多个特征了，PageRank就是其中一个特征，就考虑把机器学习的方法应用在搜索中。LTR首先通过训练数据集得到一个模型，再通过这个模型去预测新的数据。它主要分为三种，一个是基于点的Pointwise算法、一个是基于对的Pairwise算法、还有一个是基于列的Listwise算法。

面试官：那它的数据集是怎样的？怎样评价其结果？

作者：我们采用的数据集是微软提供的，共136个特征，其中第一列表示查询label，第二列对于qid查询id，紧接着是136维特征，最后是输出是一个评价等级。比如有5个等级0-4，其中4表示perfect完全相关，0表示完全不相关，就是预测结果看它属于什么等级。

面试官：你还没有自我介绍，你先做个自我介绍吧！

作者：好的！我叫XXX，来自XXX大学.....现在导师的研究方向是数据挖掘和自然语言处理相关的。我希望去到一家好的公司扎扎实实工作几年，然后回到自己的家乡贵州去当一名大学编程老师，因为父母都是老师，一直都想成为一名老师。

点评：由于这是第一次面试，自我介绍准备得也很仓促。我有几个疑问与君共勉，希望有心之人能解惑。

1.第一个是我自我介绍时忘记强调我擅长什么，做过的东西都是课程相关的，一个好的自我介绍应该如何准备？

2.第二个是我说出了自己以后的想成为一名老师，是否不应该说出来？

3.第三个就是最后忘记问面试官的姓名了，是否应该询问？

面试官：你Pointwise采用的是什么算法？请详细说说。

作者：我们采用PRank算法实现的，它是一种回归的算法。它存在一个打分函数，就是那136维特征和对应特征值的乘积加和，其结果是一个分数。然后五个等级，每个有一个阈值，通过这个得分和阈值比较可以判断其属于哪个

label等级。

面试官：你是直接通过这个得分与阈值进行比较吗？那阈值怎么确定？

作者：对了！我们需要先训练一个得出一个模型，这里就相当于先把阈值训练出来，再用这个阈值去预测。

面试官：对，先有个训练的过程，那你Pairwise采用什么算法？

作者：Pointwise是基于对的学习排序，我们采用的是RankNet算法实现的，它是基于神经网络的一个算法。它作个交叉熵，然后在梯度下降迭代直到求最优结果。

面试官：那你们那个损失函数是什么？

作者：我有些忘了这部分，损失函数好像主要让它的误差越来越小。

面试官：简单说说它们各自的优缺点。

作者：首先是基于点的学习排序，它是通过一个打分得到一个排序结果，比传统的搜索排序效果要好，但是由于它没有考虑相互之间、两两之间的关系，所以提出了基于对的方法。基于对的方法又没有考虑查询之间的顺序、同时有些查询结果差异很大，所以提出了基于列的方法。

面试官：那你们的结果是怎样呢？那个好？

作者：我们采用五个性能指标进行了评价，具体是什么我有些忘了（MAP、NDCG@5），结果是基于对和基于列的好于Pointwise。

面试官：数据集规模有多大？测试数据和训练数据规模分别多大？

作者：数据规模大的有1G左右，小的几百兆数据；同时我们要求是是在IBM的SuperVessel云平台下实现分布式运算，采用MapReduce实现的。我们一个四个同学，我和另一个同学研究算法，另外两个同学做分布式那块。

面试官：你的数据量也不大，数据集是哪来的？采用的是Python开发吗？

作者：数据集是微软研究院提供的，LTR常用的一个数据集，一个46维一个136维。我们采用的是Java实现的。

面试官：看你的简历，你是去IBM实习过吗？

作者：没有，我没去实习过。我们就是和IBM一起做了基于云平台下的一个应用，这是我们课程《高级软件工程》的大作业，IBM主要负责提供这个平台，包括Spark、MapReduce。

面试官：你们当时做这个简单来说就是IBM提供了一个云平台，但是其他公司也有很多云平台。

作者：对啊！其它公司也有，但是这是一个相当于公司和学校之间的一个学习合作吧。主要是让我们感受下分布式、LTR等前沿的知识。

面试官：你们有没有想过有的特征值的属性权重很小，把它删除的创新呢？

作者：没有考虑这个。你说这个问题确实存在，我们在做的是否，尤其是PRank算法，如果有的数据集中某个特征值很大，其他的值很想，整个结果都是被它影响。但我们当时由于时间比较紧，没做这方面的创新。

面试官：LTR现在我们已经在用了，特征值比你说的那个200多个大很多。那你们有没有遇到什么难点？

作者：哇啊，居然已经被应用了。我都不知道，只是以为还在学术中。我们当时遇到两个难点，一个是以前没有怎么学过机器学习这部分，现在需要学习这部分的算法并实现；另外一个是因为要求是基于分布式的，所以我写完算法如何与其他同学合作，让他明白算法怎么实现，转换成分布式的处理。

点评：写到这里，几个关键的东西呼之欲出。

1.面试官没有问我擅长什么语言，因为你申请的职位都会默认相关的技术和语言，很显然我这部分就是Python相关。

2.我的简历中写了8个项目，其它的包括Android、C#、MFC、PHP、C++等，但面试官只问了LTR（机器学习、自然语言处理、数据挖掘）相关，因为他只关心自己需要的这部分内容。

3.你需要非常仔细的准备简历中的非常相关部分的项目，而且是深度剖析，他会问得很详细很深，同时考虑一些创新点。

4.面试官真的很懂技术，因为你的简历会推荐到和你申请职位一致的人员给你面试。

面试官：看你的简历和博客，你做C#比较多，但是C#这块很少用在自然语言处理。

作者：以前因为本科各个方面都进行了学习，大三暑假自学了C#网络编程，后来又学了Android，我的博客主要都是根据我所做过的项目或作业写的。现在我在学Python这块做自然语言处理。

面试官：喔。Python做这块就没什么问题，那你会不会Linux下编程？

作者：我不会Linux，没研究过，最近也准备学习。

面试官：Linux基本的命名行知道吗？

作者：我没研究过，不知道。（如果知道，面试官肯定会问相关的命令，确实应该学！）

面试官：那你有没有出去实习过？

作者：我没有出去实习过，大一的时候去过大连东软学习了半个月。

面试官：看了你的简历，你做NLP这块还是非常少，没有跟着老师或去公司做过相关的吗？

作者：因为我们大一的课程比较多，需要把所有的课程上完。包括我写的博客基本都是课程项目和自学相关的内容，所以自然语言处理和数据挖掘这部分还在学习中。其实我一直都想去到企业扎扎实实学习，找个师傅跟着他学，现在是广度的学习，到了企业找个那个方向再向深度学习。

面试官：首先你说的那个挺好，但是你要知道企业需要的是创造价值的人，当然这里也有很多的学习机会，但是更多时间是让你去创造价值。包括你说你五年十年后去当老师挺好的，但是现在你能为我们做些什么？

点评：确实！企业是需要寻找为其创造价值的人才。写到这里我仿佛认识到了“现实很残酷”的道理。

1.如果有机会，你还是应该学习Linux下的命令行编程，这是一个走向吧！

2.去到一个公司实习，尤其是做你将来需要申请职位相关的项目非常重要；但我更多的看到的是实习测试、运维这些，自己斟酌吧！

3.我一直的想法都非常的单纯：“在校期间多学些东西，从广度上发展，养成良好的自学能力；将来到企业拥有这种学习能力再深入学习自己所在职位的知识，学到精通；工作5-10年之后，或者再读个博士或者直接回到贵州家乡当一名大学老师，教授编程及分享博客。”但显然这是不够的，所以接下来我也需要沉下心来深度学习一段时间。但我的梦想永远都不会被敲碎的，正如《当幸福来敲门》里面的一句话”那些没有成才的人总会说你也不能成才,有梦想就要学会保护它“。

面试官：你今年研究生大二上完，开学大三吗？

作者：我今年大一刚完，来年上大二。

面试官：看你导师是自然语言处理相关的，你简述说下分词？

作者：我只能说说我简单对它的理解吧！分词是以字分的。

面试官：分词不一定是按字分的，不同种类分法不同的。

作者：恩。比如现在我在淘宝输入框中输入“I am a student”这句英文是根据word词进行划分，分为"I"、"am"、"a"和"student"，而输入中文“我是一个学生”，它分为"学"+"生"，那么"学生"如何判断它是一个名词短语呢？

面试官：你说的是个研究这块大学生都知道，但具体怎么实现呢？怎样把"学生"连在一起呢？

作者：这部分我还没有深入的研究。

面试官：你说你自然语言处理和数据挖掘那块你擅长什么？

作者：因为这部分还在学习中，我可以说说我现在正在做的知识图谱相关的东西吗？

面试官：可以，我现在就在做知识图谱这块，那你简单说说你现在做的这个知识图谱吧？

作者：我现在在做的毕业设计就是旅游知识图谱相关的，首先是从维基百科、百度百科、互动百科、多源旅游网站中爬取旅游景点的数据，以前做的比较多的比如实体消歧。举个简单的例子吧！现在维基百科有苹果三个页面，一个指向苹果公司，一个指向苹果水果，一个指向苹果电影，如何来判断你需要的是具体哪个苹果呢？就可以通过上下文如“乔布斯”来确认它是苹果公司。而现在我先要做的是从维基百科Infobox消息盒中爬取旅游景点的信息，采用<实体,属性,属性值>三元组RDF存储，比如有三个“长城”，在把这些属性属性值进行实体对齐、信息融合，得到一个更丰富准确的信息。

面试官：你采用的是什么算法？

作者：实体对齐准备采用CURE聚类的算法实现，属性对齐采用Word2Vec计算相似距离实现。

面试官：具体的算法过程能描述下吗？

作者：因为现在才确定方案，还没有具体的深入研究具体的算法过程。

面试官：你这个只是说了数据挖掘那块，但是自然语言处理那块怎么实现的，我不知道。换句话说，如果现在你是我的一名员工了，你觉得你能做什么在这个部门？

作者：(其实想说知识图谱或搜索引擎的，我最初的想法是来到公司先学学公司一些基础东西，从来没有相关来到公司就开始为公司做什么东西。感觉都需要一个学习的过程，毕竟学校和公司之间的差别还是非常大的)

面试官：如果你有一个好的算法，创新的东西，你提出来，我们一起来研究提升现有的东西，这些都非常好。

点评：写到这里，确实认识到了自己很多不足，一方面NLP相关的分词这些基础的知识实现至少都需要了解。其

次是自己学习NLP总有一种感觉：“学习游泳，在岸上观望了很长时间，却迟迟不敢下水。”

面试官：你有什么问题吗？

作者：我有两个问题，一个是前面你提到的那个三个学习排序各自的优缺点是什么？另一个是如何实现分词的，我也想知道？

面试官：好的！首先是你的第一个问题，三个算法的优缺点，你基本都回答正确的。Pointwise就是没有考虑两两之间的关系，通过打分函数排序的，Pairwise而没有考虑它们之间的顺序。第二个问题分词，最简单的方法就是通过词库进行查找，比如“学生”，相当于一个词典，定义了很多名词短语；也可以通过训练模型来分词，那如果有个“潘长江”，你如何判断它是“潘+长江”还是人名“潘长江”呢？因为词库里很少存人名。这些都是需要考虑的。你还有什么问题吗？

作者：没有了，谢谢。

面试官：好的。

作者：非常感谢你告诉了我这么多东西，因为这是我第一次参加面试有些紧张，同时也认识了很多不足，接下来学习下Linux、进行深入的学习NLP相关知识，再如在项目中应该关注一些创新点。其实我一直都觉得面试是一个相互学习的过程，我能从你那学到很多东西，你也可能从我这学到一些知识。非常感谢！

面试官：恩，对。也谢谢你的这次面试！继续加油，挂了啊。

作者：好的，谢谢您。拜拜~

点评：你需要记住面试官会让你提问，我也没准备什么好的问题，如“职业规划”、”公司如何培养“，就问了两个他给我提的问题，至少让我知道更多。还是那句话，我认为面试是一个相互学习的过程，能让我找到很多自己的不足，更清晰的看清自己的编程方向，同时需要保持一颗平常心去对待。

四. 面试总结

终于45分钟的人生第一次面试结束了，内容基本与当时一致的，可能忘记一些细节和顺序，自己收获颇多。也希望对你有所帮助，同时是当时的所作所得所感吧！如果有错误或不足之处，还请海涵~

自己的优缺点非常明显：

优点：

1.大学期间学得比较杂，东西比较多：C++、C#、Java、PHP、Python、NLP等；
2.学习能力比较强，至少还是懂些基础的东西；
3.喜欢分享知识和博客，至少做过很多的东西，虽然比起企业项目小儿科；
4.只想找个公司，扎扎实实学习工作，没有好高骛远，且梦想是老师。

缺点：

1.没有去公司实习过，尤其是在公司做过NLP和数据挖掘相关项目；
2.不会Linux下编程，Python还没有精通；
3.自然语言处理和数据挖掘、搜索推荐相关知识没有深入学习，但又想从事这方面；
4.没有深入准备自己所做过的项目LTR和面试简历及问题；
5.学习方面仅有广度没有深度，这也不一定是坏事吧。

提升：

1.准备学学Linux下编程，鸟哥私房菜从大一就想看，五年都过去了；
2.准备想想自己需要从事哪个方向，并深入学习下该方向知识；
3.把自己的毕设知识图谱相关的做好，深入学习，因为与找工作也相关；
4.准备些基础知识，包括《编程之美》、《剑指offer》、《面试宝典》等，因为可能会有笔试；
5.如果你有机会，建议出去公司实习，建议大公司且与自己相关的部门；
6.建议你好好准备一份简历，项目最好与你所投部门相关，并且自己深入了解细节。

最后分享最近自己看的一本书里的一些内容，《你在忙什么》，放平心态，工作迟早会有，梦想却不能被敲碎。

发明机器，本想有更多的休息时间，结果人越来越累；发明手机，本想交流越来越多，结果人越来越孤独；发明网络，本想让人更有智慧，结果日益迟钝。人们猜到了开始，却猜不到结果，这是因为忽略了什么？

一部高档手机，70%的功能是没用的；一款高档轿车，70%的速度是多余的；一屋漂亮衣物，70%是闲置没空穿的；一生赚再多的钱，70%是留给别人花的。可见，累了一辈子，争了一辈子，自己能用上的也不过30%，欲望满足了也是苦。

信息爆炸的今天，各种欲望扑面而来，每个人不仅在外面不停地忙，内心也特别忙，忙了一辈子，最后却不知道自己到底在忙什么。因此，寻找让心宁静的智慧，尤为重要。束缚你的永远是你自己，所以解开它的，也只有靠你自己。

——索达吉堪布

（By:Eastmount 2015-7-13 下午6点 http://blog.csdn.net/eastmount/）