笑笑的程序人生

不以物喜，不以己悲

个人资料

ygrx

访问： 299953次
积分： 3211
等级： BLOG 5
排名： 第8675名

原创： 53篇 转载： 1篇
译文： 0篇 评论： 174条

关于我

一个已经过了而立之年的码农。
成长在湘江之畔
求学于麓山脚下
工作在帝都北京
之前一直做嵌入式方向的开发，后来慢慢转向了管理，中间有很长时间没有写过代码了，但是自己还是非常热爱代码的工作，后来，终于下决心转行到互联网做一个码农。
一个机缘巧合的机会，在一家成立了很长时间的电商网站负责整个网站的搜索和排序系统，让我过了一个非常精彩的2014年，而现在，我又有新的征程了。
C/C++,go,Python,Obj-c是我的菜
hadoop和strome也还行偶尔也吃一些erlang,haskell。
算法也能聊聊，机器学习也能聊聊，搜索引擎也能聊聊，推荐系统也能聊聊，mips和arm也能聊聊，驱动程序也能聊聊，FPGA也能聊聊，呵呵，但都仅仅是聊聊啊。。。
平时就是看书，跑步，画画

[推荐算法]基于用户的协同过滤算法

标签： 算法 python 协同过滤算法 推荐算法

2013-11-12 14:12 39036人阅读 评论(22) 收藏 举报

分类：

算法（14）

目录(?) [+]

什么是推荐算法

推荐**算法**最早在1992年就提出来了，但是火起来实际上是最近这些年的事情，因为互联网的爆发，有了更大的数据量可以供我们使用，推荐算法才有了很大的用武之地。

最开始，所以我们在网上找资料，都是进yahoo，然后分门别类的点进去，找到你想要的东西，这是一个人工过程，到后来，我们用google，直接搜索自己需要的内容，这些都可以比较精准的找到你想要的东西，但是，如果我自己都不知道自己要找什么肿么办？最典型的例子就是，如果我打开豆瓣找电影，或者我去买说，我实际上不知道我想要买什么或者看什么，这时候就可以派上用场了。

推荐算法的条件

推荐算法从92年开始，发展到现在也有20年了，当然，也出了各种各样的推荐算法，但是不管怎么样，都绕不开几个条件，这是推荐的基本条件

- 根据和你共同喜好的人来给你推荐
- 根据你喜欢的物品找出和它相似的来给你推荐
- 根据你给出的关键字来给你推荐，这实际上就退化成搜索算法了
- 根据上面的几种条件组合起来给你推荐

实际上，现有的条件就这些啦，至于怎么发挥这些条件就是八仙过海各显神通了，这么多年沉淀了一些好的算法，今天这篇文章要讲的基于用户的协同过滤算法就是其中的一个，这也是最早出现的推荐算法，并且发展到今天，基本思想没有什么变化，无非就是在处理速度上，计算相似度的算法上出现了一些差别而已。

基于用户的协同过滤算法

我们先做个词法分析 基于用户 说明这个算法是以用户为主体的算法，这种以用户为主体的算法比较强调的是社会性的属性，也就是说这类算法更加强调把**和你有相似爱好的其他的用户的物品**推

GitHub:https://github.com/wyh267
欢迎交朋友:
wyh817@gmail.com

文章分类

Direct3D学习笔记 (4)
iphone开发 (3)
linux (1)
杂谈 (5)
c/c++ (9)
技术 (22)
Javascript (1)
算法 (15)
云计算 (4)

文章存档

2016年04月 (3)
2015年06月 (1)
2013年11月 (2)
2013年10月 (3)
2013年09月 (8)
展开

阅读排行

如何在github上发起一个 (39283)
[推荐算法]基于用户的协 (39023)
文本相似度计算-Jaccard (23611)
C++多线程框架（三）--- (14463)
Go语言简单的TCP编程 (14105)
LWIP轻量级TCPIP协议栈 (13258)
Nginx一个IP配置多个主 (11906)
搭建自己的XenServer+C (11529)
搭建一个个人博客 (8840)
C++多线程框架（一）--- (7198)

评论排行

[推荐算法]基于用户的协 (22)
字符串消除 (21)
Direct3D学习笔记（三） (17)
Direct3D学习笔记（一） (15)
数组排序 --- 庞果 (12)
C++多线程框架（三）--- (9)
关于技术 (9)
Direct3D学习笔记（四） (8)
寻找直方图中面积最大的 (8)
搭建一个个人博客 (7)

最新评论

[推荐算法]基于用户的协同过滤算
野猪: reload(sys)是什么？
关于技术
飞信天下: 楼主的文章让我反思自己，什么都搞，就是什么都不精。谢谢你的文章。
Go语言简单的TCP编程
Dannyoung: 写得好仔细的
子序列的个数 --- 庞果网
菜鸟cx: 请问庞果网的 算法在哪提交

荐给你，与之对应的是基于物品的推荐算法，这种更加强调把**和你你喜欢的物品相似的物品**推荐给你。

然后就是 协同过滤 了，所谓 协同 就是大家一起帮助你啦，然后后面跟个 过滤 ，就是大家是商量过后才把结果告诉你的，不然信息量太大了。。

所以，综合起来说就是这么一个算法，那些和你有相似爱好的小伙伴们一起来商量一下，然后告诉你什么东西你会喜欢。

算法描述

相似性计算

我们尽量不使用复杂的数学公式，一是怕大家看不懂，难理解，二是我是用mac写的blog,公式不好画，太麻烦了。。

所谓计算相似度，有两个比较经典的算法

- Jaccard算法，就是交集除以并集，详细可以看看我这篇 文章。
- 余弦距离相似性算法，这个算法应用很广，一般用来计算向量间的相似度，具体公式大家google一下吧，或者看看这里
- 各种其他算法，比如欧氏距离算法等等。

不管使用Jaccard还是用余弦算法，本质上需要做的还是求两个向量的相似程度，使用哪种算法完全取决于现实情况。

我们在本文中用的是余弦距离相似性来计算两个用户之间的相似度。

与目标用户最相邻的K个用户

我们知道，在找和你兴趣爱好相似的小伙伴的时候，我们可能可以找到几百个，但是有些是好基友，但有些只是普通朋友，那么一般的，我们会定一个数K，和你最相似的K个小伙伴就是你的好基友了，他们的爱好可能和你的爱好相差不大，让他们来推荐东西给你（比如肥皂）是最好不过了。

何为和你相似呢？简单的说就是，比如你喜欢 macbook,iphone,ipad ，A小伙伴喜欢 macbook,iphone,note2小米盒子,肥皂，蜡烛 ,B小伙伴喜欢 macbook,iphone,ipad肥皂,润肤霜 ,C女神喜欢 雅诗兰黛,SK2,香奈儿 ，D屌丝喜欢 ipad,诺基亚8250,小霸王学习机 那么很明显，B小伙伴和你更加相似，而C女神完全和你不在一个档次上，那我们推荐的时候会把 肥皂 推荐给你，因为我们觉得肥皂可能最适合你。

那么，如何找出这K个基友呢？最直接的办法就是把目标用户和**数据库**中的所有用户进行比较，找出和目标用户最相似的K个用户，这就是好基友了。

这么做理论上是没什么问题的，但是当数据量巨大的时候，计算K个基友的时间将会非常长，而且你想想就知道，数据库中的大部分用户其实和你是没有什么交集的，所没必要计算所有用户了，只需要计算和你有交集的用户就行了。

要计算和你有交集的用户，就要用到`物品到用户的反查表`，什么是反查表呢？很简单，还是是上面那个AB小伙伴和C女神的例子，反查表就是喜欢macbook的有`你，A，B`，喜欢iphone的有`你，B`。。。就是喜欢某些物品的用户，有了这个表，我们就可以看出来，和你有关系的用户就只有A和B，D了，而C女神和你没有任何交集，所以不用去想C了。

这样，我们有了A和B,D，然后就分别计算A和B,D与你的相似度，不管用哪个相似性公式，我们算出来都是B和你更相似(在这个例子中，一般会用Jaccard来计算，因为这些向量不是特别好余弦化)，**但如果此时我们的`K`设定为2，那么我们就得出了与你最相邻的基友是B和A。**

这就是与目标用户最相邻的K个用户的计算。

通过这K个用户来推荐商品了

好了，你的好基友我们也算出来了，接下来要向你推荐商品了。但是我们可推荐的商品有`小米盒子，note2，蜡烛，润肤霜，肥皂`这么四种，到底哪种才是你需要的呢？这里的算法就比较广泛了，我们可以不排序，都一股脑推荐给你，但这明显可能有些你不怎么感兴趣，我们也可以做一些处理，假如我们算出来A和你的相似度是25%，B和你的相似度是80%，那么对于上面这些产品，我们的推荐度可以这么来算

- 小米盒子: 1*0.25 = 0.25
- note2: 1*0.25 = 0.25
- 蜡烛: 1*0.25 = 0.25
- 润肤霜: 1*0.8 = 0.8
- 肥皂: 1*0.8+1*0.25=1.05

这样就一目了然了，很明显，我们会首先把肥皂推荐给你，这个可能是你最需要的，其次是润肤霜，然后才是蜡烛，小米盒子和note2。

当然，你可以把上述结果归一化或者用其他你觉得合适的方式来计算推荐度，不管怎么算，推荐度还是得和基友与你相似度有关系，就是那个0.8和0.25一定要用上，不然前面白算了。

算法总结

好了，通过这个例子，你大概知道了为什么会推荐肥皂给你了吧，这就是基于用户的协同推荐算法的描述，总结起来就是这么几步

1. 计算其他用户和你的相似度，可以使用反差表忽略一部分用户
2. 根据相似度的高低找出K个与你最相似的邻居
3. 在这些邻居喜欢的物品中，根据邻居与你的远近程度算出每一件物品的推荐度
4. 根据每一件物品的推荐度高低给你推荐物品。

比如上面那个例子，首先，我们通过反查表忽略掉了C女神，然后计算出A和B,D与你的相似度，然后根据K=2找出最相似的邻居A和B，接着根据A,B与你相似度计算出每件物品的推荐度并排序，最后根据排好序的推荐度给你推荐商品。

怎么样，是不是很简单啊。

算法存在的问题

这个算法实现起来也比较简单，但是在实际应用中有时候也会有问题的。

比如一些非常流行的商品可能很多人都喜欢，这种商品推荐给你就没什么意义了，所以计算的时候需要对这种商品加一个权重或者把这种商品完全去掉也行。

再有，对于一些通用的东西，比如买书的时候的工具书，如现代汉语词典，新华字典神马的，通用性太强了，推荐也没什么必要了。

这些都是推荐系统的脏数据，如何去掉脏数据，这是数据预处理的时候事情了，这里就不多说了。

来个实战的吧

说了这么多，肥皂也推荐了，那么我们来点实际的，我这里下载了movieLens 的数据集，至于这个集合是什么大家google一下，反正很多地方用来做**测试**算法的数据，这个数据集里面有很多用户对于电影的打分，我们的需求是随便输入一个用户，然后根据协同算法，给他推荐一些个电影。

由于用户给电影打分有好有坏[1到5分]，而我们上面的例子中都是说的喜欢某件物品而没有说不喜欢的情况，所以首先，我们要把数据处理一下，简单的来做，我们可以认为3分以上的话代表这个用户喜欢这个电影，否则就是不喜欢，这样显得有点太死板了，我们也可以这么来定义，比如用户A对30部电影打分了，首先求出他打分的平均值，然后高于这个平均值的我们觉得用户喜欢这个电影，否则认为他不喜欢。

好了，用户的喜欢与不喜欢的问题解决了。下面就可以开始算法了，代码不全贴出来了，贴个流程吧，具体代码可以去看我的github

```
#读取文件数据
test_contents=readFile(file_name)
#文件数据格式化成二维数组 List[[用户id,电影id,电影评分]...]
test_rates=getRatingInformation(test_contents)
#格式化成字典数据
#     1.用户字典：dic[用户id]=[(电影id,电影评分)...]
#     2.电影用户反查表：dic[电影id]=[用户id1,用户id2...]
test_dic,test_item_to_user=createUserRankDic(test_rates)
#寻找邻居
neighbors=calcNearestNeighbor(userid,test_dic,test_item_to_user)[:k]
#计算推荐列表
recommend_dic={}
for neighbor in neighbors:
    neighbor_user_id=neighbor[1]
    movies=test_dic[neighbor_user_id]
    for movie in movies:
        if movie[0] not in recommend_dic:
            recommend_dic[movie[0]]=neighbor[0]
        else:
            recommend_dic[movie[0]]+=neighbor[0]

#建立推荐列表
recommend_list=[]
for key in recommend_dic:
    recommend_list.append([recommend_dic[key],key]
recommend_list.sort(reverse=True)
```

对于随便输入一个用户，我们得到以下这个推荐结果

```
    movie name                 release
=========================================================
Contact (1997)                 11-Jul-1997
Scream (1996)                  20-Dec-1996
Liar Liar (1997)               21-Mar-1997
Saint, The (1997)              14-Mar-1997
English Patient, The (1996)    15-Nov-1996
Titanic (1997)                 01-Jan-1997
Air Force One (1997)           01-Jan-1997
Star Wars (1977)               01-Jan-1977
Conspiracy Theory (1997)       08-Aug-1997
Toy Story (1995)               01-Jan-1995
Fargo (1996)                   14-Feb-1997
```

多输入几个用户你就会发现，像Titanic，Star Wars这种超级热门的电影，只要你选的这个用户没看过，推荐系统就一定会推荐给你，这就是我们前面说的脏数据，实际系统中这种数据是需要处理掉得。我们这篇文章只做算法讲解，就不去管这些东西了。

上一篇 杨辉三角形变型【庞果网】
下一篇 一个go语言实现的短链接服务

我的同类文章

算法（14）

- 杨辉三角形变型【庞果网】 2013-11-08 阅读 1352
- 文本相似度计算-Jaccard… 2013-10-15 阅读 23609
- 子序列的个数 --- 庞果网 2013-09-18 阅读 2949
- 数组排序 --- 庞果 2013-09-04 阅读 2469
- 不可表达的数 --- 梅森数 … 2013-08-29 阅读 1268
- 字符串消除 2013-08-28 阅读 3194
- 24点计算 --- 庞果 2013-07-30
- 寻找直方图中面积最大的… 2013-07-24

参考知识库

猜你在找

查看评论

18楼 野猪 2017-01-16 17:48发表

reload(sys)
是什么？

17楼 chengchengwoheni 2016-08-17 19:51发表

博主，求你把源码和数据集发我一份，邮箱1147841113@qq.com

16楼 Memorycollector 2016-05-20 17:46发表

博主可以跑了。请问哪里可以更改用户呢？这样每次出来结果都和你一样啊0.0

15楼 Memorycollector 2016-05-20 13:50发表

-.-楼主我跑不了啊。下载了数据集。模块也装了，不知道楼主你用的那个版本的texttable呢？

14楼 hhf457764906 2016-03-10 11:38发表

楼主你写的这篇文章真的是既幽默又浅显易懂，我刚照着这篇文章的思路在项目里面加入了推荐算法，谢谢~

13楼 听风吹雨he 2016-02-27 14:23发表

请问博主，我看到你在主程序中的recommendByUserFC邻居数量定为80，但是在定义这个函数时，参数写了K=5，有什么区别嘛？

12楼 Tesilla 2015-12-13 13:44发表

texttable下载下来安装了之后还是不能运行，这是什么原因呢？需要配置环境什么的吗？

11楼 whwstar 2015-11-17 17:51发表

楼主，能把你的源码加文件给我发一下不，非常感谢楼主，我邮箱是584855528@qq.com 我去看源码了，但是没找到文件。

10楼 baidu_28919081 2015-06-10 16:26发表

u.data能不能发一下？
515324677@qq.com
谢谢！

9楼 fengvsyou 2015-06-02 14:44发表

请问楼主能分享 Texttable 文件吗？？ 十分感谢

Re: ygrx 2015-06-16 19:23发表

回复fengvsyou：这是个python的模块，专门打印表格的，pypy上有

8楼 sinat_27154107 2015-04-05 14:51发表

博主，问下Texttable是什么

Re: ygrx 2015-06-16 19:23发表

回复sinat_27154107：这是个python的模块，专门打印表格的，pypy上有

7楼 baidu_15413957 2015-04-02 16:14发表

博主 请问你代码中“使用 |A&B|/sqrt(|A || B |)计算余弦距离”就是所谓的JaccardSimilarity方法吗

6楼 baidu_15413957 2015-04-02 16:12发表

博主 请问你代码中“使用 |A&B|/sqrt(|A || B |)计算余弦距离”就是所谓的JaccardSimilarity方法吗

5楼 wanghuirainy 2014-12-19 10:51发表

我是一个初学者，能不能把完整代码发给我

4楼 lifehack 2014-05-23 15:48发表

博主的排版很漂亮，请问是用markdown写的吗？

3楼 shellyhxl 2014-04-22 09:58发表

能留个QQ号么，关于协同过滤算法的代码问题，想跟您请教。谢谢~

2楼 暴走Q蛋 2014-04-16 08:49发表

运行结果显示ImportError: No module named texttable。想问下代码中的from texttable import Texttable是什么意思？Texttable是一个要自己下载的模块么？

Re: sinat_27154107 2015-04-05 14:31发表

回复暴走Q蛋：Texttable是什么模块？你最后解决了吗？

Re: mingzai226 2014-10-21 19:14发表

回复暴走Q蛋：怎么解决的啊

1楼 D_lejia 2014-03-09 15:26发表

我是一个初学者，对代码中from texttable import Texttable 是什么意思，试着运行代码的时候会提示我代码最后一行 print table.draw() 有错