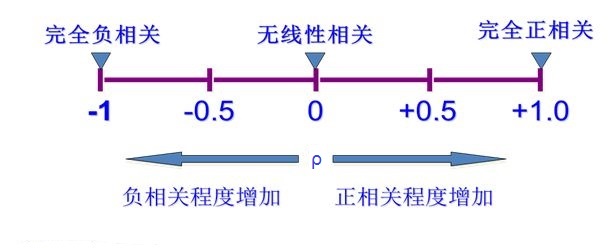
推荐系统--基于用户的协同过滤算法

一、皮尔逊相关系数计算公式



编程实现公式：

sum\_xy = 0

sum\_x = 0

sum\_y = 0

sum\_xx = 0

sum\_yy = 0

n = 0

for (x, y) in some\_condition:

n += 1

sum\_xy += x \* y

sum\_x += x

sum\_y += y

sum\_xx += pow(x, 2)

sum\_yy += pow(y, 2)

if n == 0:

return 0

#皮尔逊相关系数计算公式

denominator = sqrt(sum\_xx - pow(sum\_x, 2) / n) \* sqrt(sum\_yy - pow(sum\_y, 2) / n)

if denominator == 0:

return 0

else:

numerator = sum\_xy - (sum\_x \* sum\_y) / n

return numerator / denominator

二、数据准备

|  |  |  |
| --- | --- | --- |
| username | score | bookid |
| Liu Yi | 3 | 1001 |
| Chen Er | 4 | 1001 |
| Zhang San | 3 | 1001 |
| Li Si | 3 | 1001 |
| Liu Yi | 3 | 1002 |
| Li Si | 4 | 1002 |
| Liu Yi | 4 | 1003 |
| Zhang San | 5 | 1003 |
| Li Si | 5 | 1003 |
| Liu Yi | 4 | 1004 |
| Zhang San | 3 | 1004 |
| Liu Yi | 5 | 1005 |

三、程序

四、程序分析

1 读取user\_book.csv中的数据，放到list中

[

[‘Liu Yi’, ‘3’, ‘1001’]

[‘Chen Er’, ‘4’, ‘1001’]

[‘Zhang San’, ‘3, ‘1001’]

[‘Li Si’, ‘3’, ‘1001’]

[‘Liu Yi’, ‘3’, ‘1002’]

[‘Li Si’, ‘4’, ‘1002’]

[‘Liu Yi’, ‘4’, ‘1003’]

[‘Zhang San’, ‘5, ‘1003’]

[‘Li Si’, ‘5’, ‘1003’]

[‘Liu Yi’, ‘4’, ‘1004’]

[‘Zhang San’, ‘3’, ‘1004’]

[‘Liu Yi’, ‘5’, ‘1005’]]

]

把list中的数据，转换成dict形式。

注意，外面的dict是以用户为key，以bookid和评分构成的字典为value；里面的字典是以bookid为key，以评分为value

{

‘Liu Yi’ : {‘1001’:3.0, ‘1002’:3.0, ‘1003’:4.0, ‘1004’:4.0 ,‘1005’ 5.0},

‘Chen Er’ : {‘1001’ : 4.0},

‘Zhang San’ : {‘1001’ : 3.0, ‘1003’ : 5.0, ‘1004’ : 3.0},

‘Li Si’ : {‘1001’ : 3.0, ‘1002’ : 4.0, ‘1003’ : 5.0}

}

计算Li Si与Liu Yi的距离：

sum\_x = 12,

sum\_y = 10,

sum\_xx = 50,

sum\_yy = 34,

sum\_xy = 3\*3 + 4\*3 + 5\*4 = 41

denominator = sqrt[(50 - 144/3)(34 - 100/3)] = sqrt(12/9)

numerator = 41 - (12 \* 10)/3 = 1

distance = 0.866

计算Li Si与Chen Er的距离：

sum\_x = 3

sum\_y = 4

sum\_xx = 9

sum\_yy = 116

sum\_xy =12

denominator = sqrt[(9 - 9/1)(16 - 15/1)] = 0

distance = 0

计算Li Si与Zhang San的距离：

sum\_x = 3 + 5 = 8

sum\_y = 3 + 5 = 8

sum\_xx = 9 + 25 = 34

sum\_yy = 9 + 25 = 34

sum\_xy = 9 + 25 = 34

denominator = sqrt[(34 - 64/2)（34 - 64/2）] = 2

numerator = 34 - 64/2 = 2

distance = 1

按距离大小排序，得到的distances:

[(‘Zhang San’, 1),

(‘Liu Yi’, 0.866),

(‘Chen Er’, 0)

]

|  |  |
| --- | --- |
| Name | Weight |
| Zhang San | 1 / (1 + 0.866) = 0.536 |
| Liu Yi | 0.866 / (1 + 0.866) = 0.464 |
| Chen Er | 0 / (1 + 0.866) = 0 |

|  |  |  |
| --- | --- | --- |
| Name | BookId | recommendation |
| Zhang San | 1004 | 3 \* 0.536 = 1.608 |
| Liu Yi | 1004 | 1.608 + 4 \* 0.464 = 3.464 |
| Liu Yi | 1005 | 5 \* 0.464 = 2.32 |
| Cher Er | - | - |