基于协同过滤的用户喜好智能推荐模型

摘要

在互联网技术迅猛发展的今天，研究和建立完整的智能推荐系统有助于在海量信息中获得自己的信息和使自己的信息得到关注。本文选取 Movie Lens 网站提供的观影用户的数据，通过对数据统计筛选，结合 TOPSIS 算法建立用户喜好评价模型，结合协同过滤算法建立电影推荐模型，再综合运用这两种方法及联合聚类相关方法建立电影推荐系统。具体思想分析如下：

对于问题一，我们采用 TOPSIS 法建立了用户喜好评价模型。该问题的关键是大数据的处理和 TOPSIS 核心算法。先用 Excel 对原始数据进行筛选，得到不同用户观看电影的数量信息及评分信息，数据预处理中对多类型电影的得分采取平均分配给每一类的原则类别数对电影打分的影响。以用户的电影评分和观影次数作为评价指标，采用均方差权值法对二者加权，并用 TOPSIS 算法求解，得到编号为 108 等所给 10 个用户对 18 类不同类型电影的综合得分排序，并以此对用户喜好评价，效果良好。

对于问题二，我们依据协同过滤原理建立电影推荐模型，并从用户自身喜好和相似用户喜好两个角度出发进行推荐。首先，结合问题一，我们利用所得用户喜好矩阵和电影所属的类型成分信息，构造两个向量分别表示用户的喜好类型及电影类型，并以两者之积表示电影对于用户的迎合程度，并对之进行排名。另一方面，使用 MATLAB 以用户所属年龄组、职业为指标，筛选出用户的相似用户群，进行群个体之间的相似性计算，得到所给用户对应的相似度最大的用户的评分结果。综合以上两项因素，分别给 10 位用户推荐 5 部电影。

对于问题三，我们采用联合聚类法建立新用户智能电影推荐模型，结合问题一、二所建立的模型设计出新的智能推荐系统。该系统接受新用户的注册信息，并以联合聚类的方法和老用户的系统数据库进行聚类，得到由该新用户及其相似的老用户组成的邻居用户群，结合问题一的方法建立老用户的喜好分析模型，并以问题二的方法对这些老用户进行电影推荐，将结果作为对新注册用户的电影推荐输出。模型结构相对完整，便于推广。

关键字：智能推荐　协同过滤　TOPSIS 法　相似性分析　联合聚类

一、问题重述

在互联网技术迅猛发展的年代，如何在海量信息中找到自己感兴趣的信息，如何使自己的信息脱颖而出、受到大家的关注，是很困难的事情同时也是我们比较关心的问题。推荐系统是解决这个问题的有力工具，它通常分为收集用户信息的行为记录模块，分析用户喜好的模型分析模块和推荐算法模块。推荐系统通过建立用户和信息产品之间的关系，利用已有的选择过程或相似性关系，一方面挖掘用户潜在感兴趣的信息，另一方面让信息能够展现在对它感兴趣的用户面前。如今很多我们熟悉的各大电子商务网站都把个性化推荐作为重要的营销手段之一，推荐的类型包括商品、音乐、电影等，像 Amazon（亚马逊）、淘宝等网站很大一部分销售额来自它们的推荐系统。

依据附件所给 Movie Lens 数据集，解决下述问题。

问题 1：建立分析用户喜好的数学模型，并对编号为 108，133，228，232，336，338，545，613，696，777 的用户喜好进行分析。

问题 2：建立电影推荐的数学模型，并给问题 1 中所列用户各推荐 5 部电影（同等条件下电影编号最小者优先）。

问题 3：对于新用户（仅有用户注册信息），设计一个推荐系统。

二、问题分析

2.1 问题一的分析

问题一需要根据用户对电影的评分进行用户喜好的分析，属于评价类问题。题目中给出了用户所观看的电影的一些信息，通过分析可知，用户观看同一类别电影的次数和对不同电影的评分均会反映用户的喜好类型，故选用这两个因素作为评价指标。由于同一个电影可能隶属于不同的电影类别，所以相同的评分结果参考价值可能不同，我们题目中对评分结果进行加权，权值取该电影所属类别总数的倒数。将加权后的评分结果进行累加，采用取平均值的方法剔除评分中“次数”的影响，从而得到一个用户对不同类型电影的 1*18 的评分矩阵。将数据导入 matlab，通过编程统计出这 10 个用户对不同类型电影观看次数的 1*18 的频数统计矩阵。对这两个矩阵数据进行归一化处理，采用均方差权值法对这两个指标进行加权。接着采用 TOPSIS[1]（理想解法）得出每一个用户对不同类型电影的综合得分并进行排名，根据排名结果对该 10 个用户进行喜好分析。

故我们建立一个基于 TOPSIS 法的用户喜好评价模型，并对题目中所给用户进行喜好评价。

2.2 问题二的分析

问题二要求建立电影推荐的模型，我们从用户历史评价信息及其所在群体的选择信息两方面入手，即给用户推荐的影片可以从他的喜好和他所在的相似性群体出发生成。参考第一问得到的用户喜好分析模型对电影进行打分，得到从用户喜好角度出发的推荐影片排序；并以年龄和职业为依据，筛选数据并进行相似性

分析，找出与用户相近的群体，得到该群体的喜好，并此来预测用户未看过的电影的虚拟评分，得到用户可能会喜欢的影片排序，最后在两种指标排序中通过选取适当权值，以得分最高且用户未看过的电影进行推荐。

故我们建立一个基于用户协同过滤[2]的电影推荐模型，并对题目中所给每一位用户推荐五部电影。

2.3 问题三的分析

该问题是要对新注册用户的属性特征进行分析，根据已搜集到的信息对新用户的兴趣进行分析，推荐他可能感兴趣的电影。由于新用户没有历史浏览记录，无法根据他观看影片的种类，情况和评分分析出他的兴趣爱好，所以要给他推荐电影，可以先运用联合聚类法[3]找出与他年龄，职业，性别，学历等属性相近的用户，在系统数据库调用其电影评分等数据，根据问题一所建立的模型分析相似用户的喜好类型。再通过问题二建立的电影推荐模型所推荐的电影作为新用户的推荐输出。

故建立一个基于联合聚类法的新用户推荐模型，并结合问题一二所建立的模型设计出功能完整的智能推荐系统。该系统接受用户的注册信息并根据数据库相关信息给出推荐电影作为输出。

三、模型假设

1. 假设所有统计的数据客观可靠，符合实际。
2. 假设每个用户对电影的评价只与其喜爱程度有关，不受其他因素的影响。若用户对电影的评价受亲人朋友的影响，他的评价就不具有代表意义，作为数据来统计时容易出现误差。
3. 假设所有用户对电影的评价标准相同。若标准不一样，则将这些用户对电影的评分统计并分析是没有多少意义的。
4. 假设用户对电影的喜爱程度只与电影的类型相关。若每部电影的好坏对用户的评价影响较大，则无法根据用户所看电影和评分来判断他所喜欢的类别。
5. 假设与某一用户相似的个体主要分布在与其同职业、年龄相近的群体之中。这样会使我们在问题的求解时，对推荐用户找的邻居更为合理，使得推荐的电影更有可能被用户喜欢。
6. 假设每个年龄区间内的用户兴趣爱好相近，不同年龄区间内的用户兴趣爱好相差较远。这样我们可以不用进行年龄的模糊分析，直接将用户年龄分成几个具体区间，使得更有利于聚类和协同过滤分析。
7. 假设用户看某一部电影时，对该电影所属的所有类别喜好程度一致。

四、符号说明

符号	意义
x_{ij}	第 i 部电影对应属性类别
p_i	用户对第 i 部电影的打分
Q_i	第 i 部电影打分向量
q_{ij}	该电影单个分量的加权得分
y_j	用户对第 j 类电影的喜好程度
STD_j	第 j 个指标的标准化值
W_i	第 i 个指标的权重
G_i	第 i 个用户对电影的喜好程度
g_{ij}	第 i 个用户对第 j 类电影的综合打分
H_J	第 j 部电影的特征代表向量
$h_{j,n}$	第 j 部电影所属第 n 个类别
L_{ij}	第 i 个用户对第 j 部电影的喜好程度
r_{ij}	用户 i 对项目 j 的评价值
$sim_{ua,ub}$	用户 a、b 之间的相似度
$\overline{r_i}$	用户 i 对所有项目评价的均值
n_1	两个用户同时选择电影的数目
n_2	两个用户评价过得电影总数目

五、模型的建立与求解

5.1 问题一的分析和求解

5.1.1 数据的处理

附件给出了 943 名用户对 1682 部影片的评分情况、及影片所属类别。数据中编号为 267、1373 的两部电影由于不属于任何电影类别而被剔除，故题目的有

效电影数据为 1680.用 Excel 分别筛选出 10 位用户所观看电影的评分。

题目给出了 18 种不同的电影分类类型，结合附件的原始数据可知，同一个电影可能会属于多个不同的电影类型而导致其参考价值不同，因此不能简单地将此评分结果进行计算。选用某一部电影所属类别的倒数作为该电影评分的权值，最后通过加权求和的方式对数据进行第一步处理，同时将每个用户的给予一个 1*18 的向量来描述他对每一类的喜好。具体处理过程如下:

1、筛选出某一个用户所有看过的 n 部电影和第 i 部电影对应所属类别的数据 x_{ij} 以及他对每部电影的打分 p_i，其中：

$$x_{ij}=\begin{cases}1, & \text{第i部电影属于第j个类别；}\\ 0, & \text{第i部电影不属于第j个类别。}\end{cases}$$

2、考虑每部电影所属类别数越大，其评价受到类型间组合的影响就越大，则它的对于一个类别来说参考价值就越小，我们可以定权重使得一个影片所属类别越多其分值对每一项的参考价值就越低，这里运用较为简单的倒数法，即每一部电影的得分乘以其所包含的类别数的倒数，即得到该部电影的加权打分向量 Q_i，其中的单个分量可以表示为：

$$q_{ij}=\frac{p_i.x_{ij}}{\sum_{j=1}^{18}x_{ij}} \qquad (1)$$

3、由于选用了评分结果和观影次数作为评价指标，故在按列求和得到每部影片对应的加权得分后，需要剔除得分中次数因素的影响，将得分除以每一个类别被打分的次数：

$$y_j=\frac{\sum_{i=1}^{n}q_{ij}}{\sum_{i=1}^{n}x_{ij}} \qquad (2)$$

即得到该用户对每一类的喜好程度（对于没有包含到的类别即认为用户目前对这些类别不感兴趣，之后的模型会分析其潜在感兴趣的类别，将在下文进行讨论），结果表示为一个 1*18 的向量，其分量 y_j 的大小即表示用户对每一类由打分所反映的喜好程度。此过程需将数据导入 matlab 编程实现。

用 excel 分别对每个用户所看电影进行统计，得到所属每一电影类别的数量（后记用户对该种类型的观影次数）。得到一个 1*18 的频数统计矩阵结果如下：

表 1 用户电影评价结果加权分析表

编号 / 类型	108	133	228	232	336	338	545	613	696	777
1	1.04	0.90	1.50	1.12	1.16	1.58	1.35	1.21	0.97	1.15
2	0.85	1.14	1.92	1.04	1.10	1.38	1.17	2.33	2.00	1.67
3	1.33	0.00	0.00	1.33	1.00	1.79	1.05	1.33	0.00	1.33
4	1.42	1.21	2.50	1.21	0.87	0.75	1.09	1.33	0.00	1.33
5	2.34	1.96	1.11	1.71	2.08	2.34	1.71	2.00	1.33	2.71
6	1.33	1.13	0.67	1.65	1.33	1.44	1.09	1.83	1.21	1.13
7	0.00	0.00	0.00	3.67	0.00	0.00	0.00	0.00	0.00	0.00
8	2.13	2.07	2.05	2.36	2.16	2.56	1.65	3.27	2.72	2.57
9	0.00	1.33	0.00	1.00	1.00	0.00	1.09	0.00	0.00	0.00
10	0.00	0.00	0.00	1.25	0.00	2.79	1.50	2.50	1.25	0.00
11	0.00	0.67	2.00	1.50	1.17	3.25	2.11	2.00	1.63	2.00
12	1.13	0.00	0.00	1.57	0.71	1.25	1.34	0.00	0.00	1.00
13	1.50	0.69	0.33	2.31	0.75	1.99	1.11	2.50	1.47	0.75
14	1.46	1.24	1.17	1.63	1.44	1.80	1.31	1.90	1.23	1.71
15	1.05	1.07	1.50	1.72	1.19	2.19	1.32	1.81	0.00	1.63
16	1.11	0.96	1.33	1.68	1.36	1.88	1.42	1.85	1.21	1.33
17	1.26	0.89	1.83	1.31	1.07	1.67	1.24	1.56	1.79	1.77
18	0.00	0.00	0.00	2.17	1.17	1.33	1.54	1.67	0.00	0.00

表 2 用户观影次数分析表

编号 / 类型	108	133	228	232	336	338	545	613	696	777
1	11	8	2	17	17	4	81	6	5	4
2	7	3	2	9	9	4	50	3	2	1
3	1	0	0	1	1	4	12	1	0	1
4	2	4	1	6	5	2	19	1	0	1
5	9	7	3	24	100	32	47	5	3	14
6	2	2	2	6	8	3	8	4	4	4
7	0	0	0	3	0	0	0	0	0	0
8	14	11	16	58	24	30	34	14	23	23
9	0	1	0	1	1	0	3	0	0	0
10	0	0	0	1	0	4	1	1	1	0
11	0	1	1	1	3	2	15	1	2	1
12	2	0	0	8	4	2	13	0	0	1
13	2	4	1	4	2	8	3	1	5	1
14	11	6	5	26	29	22	25	5	5	4
15	7	5	1	11	6	4	36	4	0	2
16	6	9	3	9	12	12	33	5	7	7
17	6	3	3	16	2	9	20	4	4	7
18	0	0	0	2	3	1	8	1	0	0

5.1.2 模型相关原理

均方差权值法

均方差均值法[4]的基本原理是:若指标 J 对所有的样本反之,如果均无差别,则对其样本排序将不起作用,此时可令其权值系数为 0;指标 J 能使所有样本的属性值有较大差异,这样的评价指标对样本的排序将起重要作用,此时应该给以较大的权值。即各指标权重系数的大小取决于各指标样本属性值得相对离散程度。离散程度越大,则该指标的权重系数就越大,反之权重系数就越小。我们用均方差表征标准化后的各指标属性值得离散程度,将这些均方差归一化,其结果即为各指标的权重系数。

操作过程和步骤

(1) 原始数据标准化。为了消除不同量纲造成的不可比性,采用极差标准化方法对原始数据进行标准化处理,得到各指标的标准化值(STD 值)

$$STD_j = (x_j - \min x_j) / (\max x_j - \min x_j) \quad (3)$$

此式中,STD_j 为指标 J 的标准化值,x_j 为指标 J 的原始数值,$\max x_j$ 和 $\min x_j$ 分别为指标 J 的最大值和最小值。

(2) 计算指标权重。首先计算各指标 STD 值得标准差,然后带入公式即可算出权值

$$W_j = \frac{\sigma(STD_j)}{\sum_{j=1}^{n}\sigma(STD_j)} \quad (4)$$

此式中,$\sigma(STD_j)$ 为 STD 的标准差,W_j 为第 j 个指标的权重。

TOPSIS 法相关原理

TOPSIS 法是多目标决策[5]分析中一种常用的有效方法,其基本原理就是通过比较平价对象与最优解、最劣解的距离来进行排序,若评价对象最开进最优解同时有远离最劣解,则为最好;否则为最差。该方法基于系统工程中的乘法原则,要求各项指标尽可能取得较好的水平。最优解的个指标值都达到各评价指标的最优值,最劣值的各指标值都达到各评价指标的最差值。

5.1.3 模型一的建立

模型一建立了基于 TOPSIS 法的用户喜好的评价模型,具体步骤如下:

a. 建立决策规范矩阵。

在数据预处理中,分别得到了每一位用户的电影评分及观影次数的矩阵 q_i 、n_i $i=1,2,...,18$.现以一个用户为例进行分析。建立以电影评分和观影次数为指标的决策矩阵

$$A=(a_{ij})_{2\times 18} \qquad i=1,2;\ j=1,2,\cdots,18.$$

利用标准 0-1 变换对决策矩阵进行规范化处理：

$$b_{ij}=\frac{a_{ij}-a_j^{\min}}{a_j^{\max}-a_j^{\min}},i=1,2;\ j=1,2,...,18.$$

b．构成加权规范矩阵 $C=(c_{ij})_{2\times 18}$ 各权重值的确定利用均方差权值法，得到各属性的权重向量为 $w=[w_1,w_2]^T$，其中

$$W_i=\frac{\sigma(STD_i)}{\sum_{i=1}^{2}\sigma(STD_i)},\ i=1,2.$$

故有：

$$C_{ij}=w_i^*\ b_{ij}\ i=1,2;j=1,2 \qquad (5)$$

c．确定正理想解 C^* 和负理想解 C^0。该模型中选取正理想解为（1,1），负理想解（0,0）。

d．计算各方案到正理想解与负理想解的距离。备选方案 d_i 到正理想解得距离为：

$$s_j^0=\sqrt{\sum_{i=1}^{2}(c_{ij}-c_i^0)^2},\ j=1,2,...,18; \qquad (6)$$

备选方案 d_i 到正理想解得距离为：

$$s_j^0=\sqrt{\sum_{i=1}^{2}(c_{ij}-c_i^*)^2},\ j=1,2,...,18; \qquad (7)$$

e. 计算各方案的的排队指标值 f_i^* 并由大到小排列，其中

$$f_j^*=j\,s_j^0/(s_j^0+s_j^*),\ j=1,2,...,18. \qquad (8)$$

5.1.4 模型的求解

通过 matlab 编程求解理想解法所建立模型的解，具体包括数据的归一化处理、不同用户不同权值的确立和理想解法的求解。分别得到 10 个用户对 18 类电影喜好程度的综合评分，结果如下：

表 3 用户喜好程度综合评分表

编号 / 类型	108	133	228	232	336	338	545	613	696	777
1	0.30	0.29	0.26	0.05	0.17	0.19	0.41	0.20	0.16	0.18
2	0.21	0.22	0.32	0.02	0.15	0.17	0.29	0.32	0.29	0.24
3	0.21	0.00	0.00	0.07	0.13	0.22	0.19	0.19	0.00	0.19
4	0.23	0.25	0.39	0.05	0.12	0.09	0.20	0.19	0.00	0.19
5	0.43	0.41	0.20	0.16	0.49	0.42	0.35	0.29	0.20	0.43
6	0.22	0.21	0.12	0.14	0.18	0.18	0.19	0.27	0.19	0.18
7	0.00	0.00	0.00	0.44	0.00	0.00	0.00	0.00	0.00	0.00
8	0.47	0.50	0.44	0.31	0.30	0.43	0.31	0.50	0.50	0.48
9	0.00	0.23	0.00	0.00	0.13	0.00	0.18	0.00	0.00	0.00
10	0.00	0.00	0.00	0.05	0.00	0.32	0.24	0.33	0.19	0.00
11	0.00	0.12	0.32	0.10	0.16	0.35	0.34	0.28	0.24	0.28
12	0.19	0.00	0.00	0.12	0.10	0.15	0.23	0.00	0.00	0.15
13	0.24	0.17	0.06	0.26	0.10	0.25	0.19	0.33	0.23	0.11
14	0.35	0.29	0.22	0.15	0.23	0.31	0.25	0.28	0.20	0.25
15	0.24	0.25	0.25	0.15	0.16	0.26	0.27	0.27	0.00	0.24
16	0.23	0.32	0.24	0.14	0.19	0.26	0.28	0.28	0.20	0.22
17	0.25	0.18	0.31	0.07	0.14	0.22	0.23	0.23	0.27	0.28
18	0.00	0.00	0.00	0.23	0.16	0.16	0.25	0.23	0.00	0.00

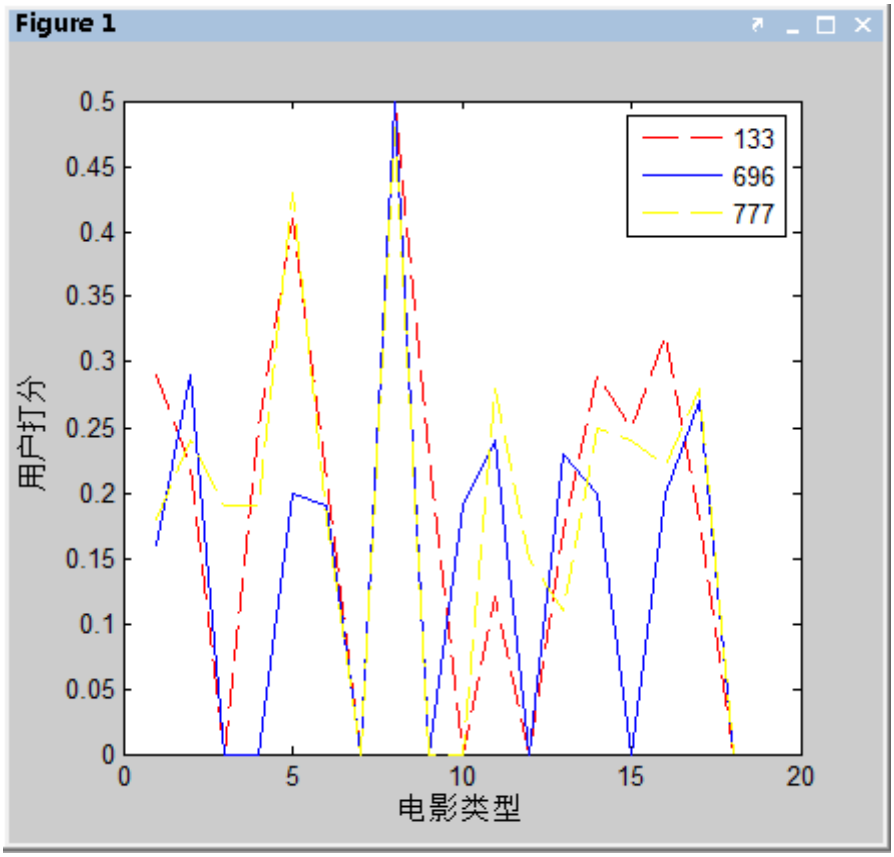


图 1 用户 133、696、777 对不同电影类别的综合评分

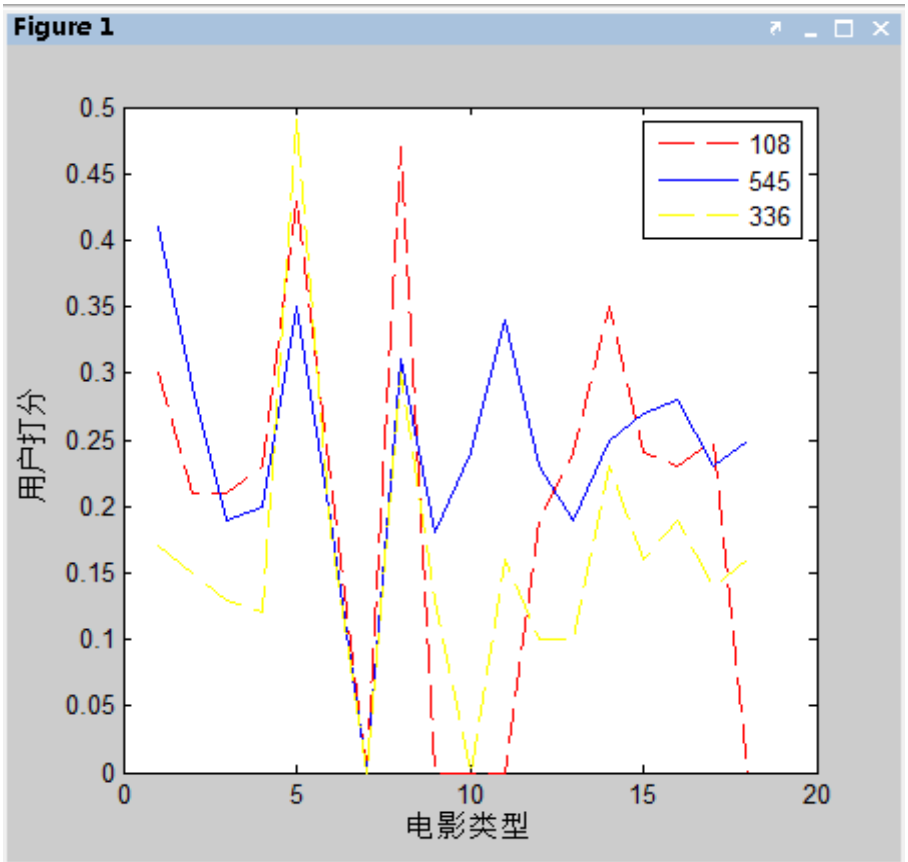


图 2 用户 108、545、336 对不同电影类别的综合评分

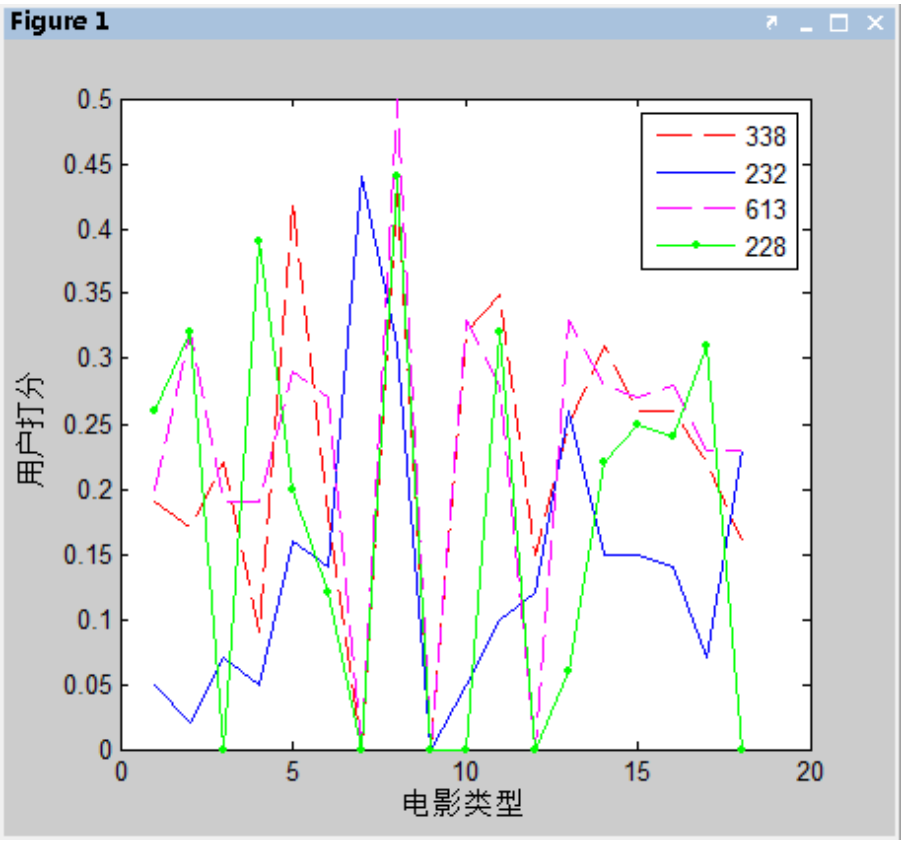


图 3 用户 338、232、613、228 对不同电影类别的综合评分

5.1.5 模型结果分析

由以上数据结果分析可以看出

Drama（剧情类）电影在 7 个用户中排名均为第一，而所给 10 个用户隶属 10 个不同职业，年龄变化范围 21—63 岁。可见剧情类电影在各类人群中均比较受欢迎；其次，喜剧和浪漫类电影的得分也普遍较高，这符合我们的认知。

对于编号 108 用户，作为教育工作者，他最喜欢的两种电影类型是剧情类和喜剧，其次是浪漫类电影及动作类电影，他不喜欢幻想类、恐怖片及黑色电影。

作为工程师的 133 用户，偏爱剧情类和喜剧、动作类电影，不喜欢纪录片、西部片，这可能与其职业特点有关及该用户本身不太喜欢看电影有关（可从其观影次数看出）。

编号为 228 的用户是名学生，该用户喜好特征比较明显。最喜欢剧情类电影，偏爱儿童剧。由于观影次数较多，故对于得分为 0 的幻想类、黑色电影、西部片等，有理由认为他不喜欢。

用户 232 较为与众不同，他最喜欢纪录片，与他科学家的职业有很大相关性；作为销售员的 336 号，可能是职业特点要求，最爱喜剧；用户 545 爱好比较广泛，最爱喜剧，对剧情片和惊悚片也有很大的爱好。

用户 338、545、613、696 用户来说，喜好电影类型相似，对剧情片、喜剧片都有明显的爱好，其他电影的喜好程度一般，且大多集中在动作冒险、恐怖惊悚等类型的电影。

5.2 问题二的分析和求解

5.2.1 模型原理介绍

协同性原理：

协同过滤式推荐系统[5]中广泛使用的最成功的推荐技术。其原理是首先根据与用户的偏好一致的邻居用户，然后对邻居用户进行分析，把邻居用户喜欢的项目推荐给目标用户。

传统的协同过滤分为 3 个阶段：(1)用户偏好的分析与表示；(2)邻居用户筛选与形成；(3)用户推荐的生成。

用户的偏好信息可用一个‘用户—项目’的评分矩阵 R 来表示， R 是一个 $m*n$ 阶的矩阵，m 表示用户数，n 表示项目数，矩阵中的每一元素 R_{ij} 表示第 i 个用户对第 j 个项目的评分值，传统的协同过滤方法在 R 上计算目标用户与其他用户间的评分相似性，选择相似性最高的前 k 个用户组成目标用户的最近邻集合。

最近邻选择[6]是协同过滤中最重要的步骤，传统的协同过滤算法在两个用户都有评分的项目集合上计算用户间的相似性，筛选最近邻集合的元素。

5.2.2 模型的建立

建立了基于协同过滤的用户电影推荐模型。

给用户（老用户）推荐电影需要考虑两方面的因素：用户历史选择情况和邻居用户的选择情况。综合这两方面来对他进行影片喜好分析与推荐。

（1）用户历史选择分析

首先，参考用户的历史选择情况我们采用了第一问得到的 10 个用户（108，133，228，232，336，338，545，613，696，777）的喜好程度综合评分表，并抽取每个用户对应的一列综合评分值作为一个用户喜好程度向量：

$$G_i = \{g_{i,1}, g_{i,2}, ..., g_{i,18}\}, (i = 1, 2, ..., 10)$$

向量的每一个元素 g_{ij} 表示第 i 个用户对第 j 个类的综合打分。由于这样得到的向量每一类电影对应的值的大小表示了用户对该类别的喜好程度，因此不论喜好程度大小对应的坐标都有一定的值，最终计算加和时，就会导致占类别数目大的电影占有一定的优势，因此需要对初步计算得到的喜好程度向量进行处理，避免由于电影所占类别较大而引起的最终喜好度的增加。这里我们采用数据过滤的方法，即选取每个向量的前五个最大值，其他的分量取为 0（若大小排名第五的分量有多个则全部选取），以此突出喜好程度大的类别的影响，消除由喜好程度小的分量和电影所占类别数带来的影响，得到处理后的喜好程度向量：

$$G'_i = \{g_{i,1}, g_{i,2}, ..., g_{i,18}\}, (i = 1, 2, ..., 10)$$

其次，提取 u_item.xls 中每个电影的一列表示所属类别的数据生成每部电影的特征代表向量：

$$H_j = \{h_{j,1}, h_{j,2}, ..., h_{j,18}\}^T \ (j = 1, 2, ...1682)$$

其中，267 和 1373 两个不予考虑。

接着，用每个用户喜好程度向量分别右乘以每部电影的特征代表向量，得到每个用户对所有电影的喜好度：

$$L_{ij} = G'_i * H_j, (i = 1, 2, ..., 10, \ j = 1, 2, ..., 1682)$$

得到每个用户对每部电影的喜好度 L_{ij}，再对 L_{ij} 按用户的喜好度排序，得到第一个用户的推荐电影序列。

（2）基于用户相似度的协同过滤模型

先对用户群体（943 个个体）进行分类筛选，步骤如下：

a．对于一个用户，先利用职业指标筛选出与之同职业的一类用户；

b．再在其中利用年龄段指标截取与之年龄相近的更小的一个集群；

c．在其中通过相似度分析找出 1~2 个邻居用户，选择他们打分较高的电影推作为预推荐序列，同时，在影片的选择上需要滤除该用户已经看过的电影。

在筛选出来的用户中使用相似度分析，需要用到两个数据集：历史评价矩阵和人口统计信息。由每个用户观看的电影和打分数据我们建立“用户—项目”评价矩阵如下：

$$R = \{r_{ij} \mid i = 1...m, \ j = 1...n\},$$

其中，m 表示用户，n 表示项目数，r_{ij} 表示用户 i 对项目 j 的评价值。用户之间的相似度可通过矩阵的行向量计算得到，为了较少用户评价的个人差异，采用改进的 $\cos ine$ 相似性来计算用户之间的相似度[7]，如下所示：

$$sim_{ua,ub}=\frac{n_1}{n_2}\frac{\sum_k\left|r_{ak}-\overline{r_a}\right|*\left|r_{bk}-\overline{r_b}\right|}{\sqrt{\sum_k(r_{ak}-\overline{r_a})^2}*\sqrt{\sum_k(r_{bk}-\overline{r_b})^2}},\qquad(9)$$

其中，r_{ak} 和 r_{bk} 分别表示用户 a 和 b 对项目 k 的评价值，r_a 和 r_b 分别表示用户 a 和 b 对所有项目的评价均值，k 为用户 a 和 b 共同评价的项目数，n_1 表示两个用户同时选择电影的数目，n_2 表示两个用户总的评价过的电影的数目。以此作为衡量两个用户关联程度的标准。

（3）关于同类人群筛选方法的说明

传统的协同过滤方法需要对大量用户的属性和喜好进行聚类，通过聚类进行数据的降维处理，得到与用户相似性较高的一个群体。此模型中，出于三方面考虑：

a. 用户信息较少，只有两个指标：年龄和职业；

b. 给某个用户推荐的电影需要参考个人喜好和相似性群体喜好两方面的因素，且个人因素在推荐时参考价值较大；

c. 与用户相似性较大的邻居用户出现在与他同职业和年龄相近的群体的概率较大。

因此我们直接使用年龄段和职业作为筛选群体的两个条件，通过筛选之后样本缩小为和用户具有相同职业且年龄相近的一个相似性群体，再在这个群体中进行用户之间的相似性分析，从而得出与用于相似度最高的邻居用户，参考其选择给用户推荐电影。

在选定年龄段划分时我们选择：0~15 岁为一段，以后每 5 岁为一个年龄段进行划分。在划分过程中遇到所得样本较少的情况，我们对其进行扩大年龄段处理，原则有两个：1、就近添加；2、最少样本数为 5，达到即停止扩大。

5.2.3 模型的求解

利用第一问的得到的数据并加以处理，将数据导入 Matlab 进行求解，得到 10 个用户处理后的喜好程度向量列表：

表 4 处理后的用户喜好程度向量列表

电影类型 / 用户编号	1	2	3	4	5	6	7	8	9
108	0.00	0.00	0.00	1.42	2.34	0.00	0.00	2.13	0.00
133	0.00	0.00	0.00	1.21	1.96	0.00	0.00	2.07	1.33
228	0.00	1.92	0.00	2.50	0.00	0.00	0.00	2.05	0.00
232	0.00	0.00	0.00	0.00	0.00	0.00	3.67	2.36	0.00
336	0.00	0.00	0.00	0.00	2.08	1.33	0.00	2.16	0.00
338	0.00	0.00	0.00	0.00	2.34	0.00	0.00	2.56	0.00
545	0.00	0.00	0.00	0.00	1.71	0.00	0.00	1.65	0.00
613	0.00	2.33	0.00	0.00	2.00	0.00	0.00	3.27	0.00
696	0.00	2.00	0.00	0.00	0.00	0.00	0.00	2.72	0.00
777	0.00	0.00	0.00	0.00	2.71	0.00	0.00	2.57	0.00
电影类型 / 用户编号	10	11	12	13	14	15	16	17	18
108	0.00	0.00	0.00	1.50	1.46	0.00	0.00	0.00	0.00
133	0.00	0.00	0.00	0.00	1.24	0.00	0.00	0.00	0.00
228	0.00	2.00	0.00	0.00	0.00	0.00	0.00	1.83	0.00
232	0.00	0.00	0.00	2.31	0.00	1.72	0.00	0.00	2.17
336	0.00	0.00	0.00	0.00	1.44	0.00	1.36	0.00	0.00
338	2.79	3.25	0.00	0.00	0.00	2.19	0.00	0.00	0.00
545	1.50	2.11	0.00	0.00	0.00	0.00	0.00	0.00	1.54
613	2.50	2.00	0.00	2.50	0.00	0.00	0.00	0.00	0.00
696	0.00	1.63	0.00	1.47	0.00	0.00	0.00	1.79	0.00
777	0.00	2.00	0.00	0.00	1.71	0.00	0.00	1.77	0.00

从表中我们可知具有 8 号（drama）、5 号(comedy)和 14 号(romance)类别的影片受到很大一部分用户的欢迎，受众面较广，这符合我们的常识。

用 matlab 编程将处理后的数据与电影所属类别构成的 0-1 矩阵求积并进行排名，得到不同各用户电影喜好度排名得到如下：

表 5 各用户电影喜好度排名列表

用户编号		108	133	228	232	336	338	545	613	696	777
电影喜好排名	1	312	560	35	135	55	184	184	312	172	944
	2	170	170	78	847	129	559	17	655	655	1204
	3	512	512	132	1366	170	561	396	184	97	170
	4	517	517	457	1561	512	569	201	201	720	512
	5	692	692	1608	312	517	784	123	299	403	517
	6	731	731	172	855	692	928	208	642	35	692
	7	775	775	8	1129	731	17	564	135	78	731
	8	778	778	842	1212	775	396	637	855	132	775
	9	936	936	423	914	778	201	853	914	457	778
	10	1100	1100	308	55	936	123	854	1129	1608	936

根据用户间相似性计算得到相应推荐影片结果如下：

表 6 邻居用户推荐电影列表

参考用户个人喜好和相似度分析结果后，我们以 4:1 的比例（比例参考用户

用户编号	108	邻居用户	133	邻居用户	228	邻居用户	232	邻居用户	336	邻居用户
电影排序	258	192	1	157	272	929	47	195	288	925
	284	192	100	157	286	929	100	195	5	925
	302	192	117	157	313	929	127	195	98	925
	1405	192	147	157	690	929	132	195	185	925
用户编号	338	邻居用户	545	邻居用户	613	邻居用户	696	邻居用户	777	邻居用户
电影排序	9	667	1	77	127	744	9	136	9	694
	28	667	28	77	302	744	14	136	22	694
	86	667	42	77	479	744	100	136	50	694
	124	667	52	77	508	744	116	136	69	694

个人喜好的参考性和相关性分析结果的可靠性）在两类电影集合里面选取推荐影片，同时，进行一次过滤，删去用户已经看过的影片，得到 10 位用户的推荐影片结果如下：

表 7 用户推荐电影的最终结果

用户编号		108	133	228	232	336
推荐电影	1	312	560	35	135	55
	2	170	170	78	847	129
	3	512	512	132	1366	170
	4	517	517	457	1561	512
	5	258	1	272	47	5
用户编号		338	545	613	696	777
推荐电影	1	184	184	312	172	944
	2	559	396	655	655	1204
	3	561	201	184	97	170
	4	569	123	201	720	512
	5	9	42	302	14	22

5.2.4 结果分析

（1）从表中所给结果分析可知，总体推荐电影的差别比较明显，由附件资料得知，这十位用户来自 10 个不同的职业，年龄分布近似正态函数，使得个体与个体之间具有一定差异，推荐影片有差异，结果合理；

（2）有两个用户（108、133）被推荐的影片相似，通过查原始数据可知，两人的职业相似，选择的电影类别及打分情况相近，故两者被推荐影片相近；

（3）170 号影片被推荐了 4 次，512 号影片被推荐了 3 次，查阅原始数据可知，两部影片在类别上可归于 8（drama）、5（comedy）和 14（romance），由用户喜好程度列表可知这两部电影较受欢迎。

5.3 问题三的分析和求解

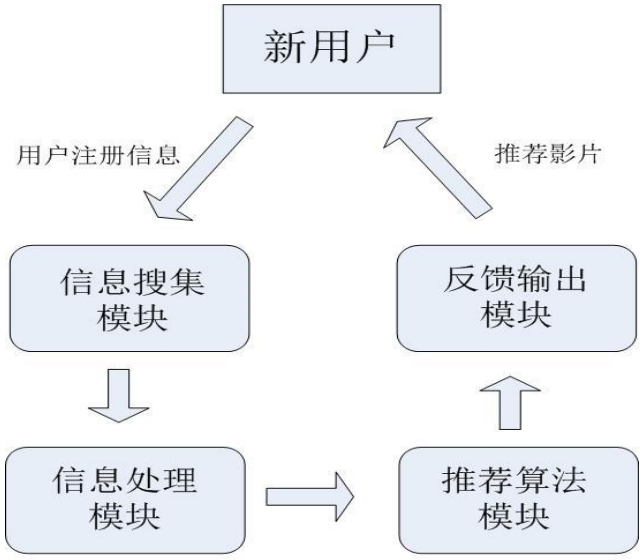
5.3.1 系统相关信息介绍

（1）推荐系统功能介绍

推荐系统[8]是利用电子商务网站模拟销售人员向客户提供各类影片的信息和建议，帮助用户选择决策观看什么影片的过程。该系统是要根据用户的注册信息建立用户模型，将有观看历史记录的用户模型的特征与新注册用户模型特征信息进行匹配，使用适合的推荐算法进行筛选过滤，先选出和新用户特征相似的用户，再挑选出用户可能会喜欢的电影并推荐给该用户。该系统从本质上来说，它在信息系统的分类中也属于决策支持系统的一个分支。决策支持系统是基于数据模型和知识库，通过对系统与用户之间的人机交互模型进行分析，辅助用户进行决策的计算机应用系统。它是信息管理系更深入层次发展而提出的一种更为智能的信息管理系统。

（2）推荐系统模块分析

如下图所示，系统主要包括用户注册信息收集模块、用户信息处理模块和推荐算法模块。具体来说，系统包括以下几个功能模块：



推荐系统流程

图 4 智能推荐系统模块示意图

a. 行为记录模块

用户模型的输入数据是用户建模的基础，它代表的是用户的注册信息，用户的注册信息包括用户的姓名、性别、年龄、所在地、学历、从事行业及职位等，是用户的基本信息。

b. 模型分析模块

用户兴趣模型的表示方法是用户兴趣模型的核心。它的形成过程是：当用户输入注册信息完毕后，服务器数据采集模块即记录下用户的每一个信息，如用户的姓名、性别、年龄、所在地、学历、从事行业及职位等，记录完毕后，得到m个特征属性值。

调出数据库中n个有观看历史记录的用户注册信息对应的$n\times m$个特征属性值，分别算出每个老用户相对新用户的属性比较值，通过聚类的方法，作相似度分析，找出属性比较值最小的用户，即是找出与该用户特征属性值最接近的用户。推荐算法模块。

c. 推荐算法模块

将前两个模块得到的数据运用协同过滤法和联合聚类法使得老用户与新用户的兴趣爱好能够合理对应并生成为新用户呈现推荐电影目录。其中协同过滤算法在第二问中已经提到，而聚类分析法是数据挖掘的一项重要的方法，它实际上是将一组数据分成若干个组，每个组里对象具有很大的相似性，不同组之间存在尽量大的差异性，再在这些组之间寻找数据之间的内在关系。

5.3.2 数据预处理

本系统的数据库中记录的人口统计信息包括用户的姓名、性别、年龄、所在地、学历、从事行业及职位等属性。这些属性都比较抽象，无法直接用，因此要对各个属性做相应的处理：将“年龄”属性做离散化处理，将“学历”“职业”、“性别”等属性符号化。

老用户有 n 个，则新用户是第 $n+1$ 个，每个用户的属性值有 m 个。

第 l 个用户的第 i 个统计值用 f_{li} 表示，即

$$F=\begin{pmatrix} f_{11} & \cdots & f_{1m} \\ \vdots & \ddots & \vdots \\ f_{n+11} & \cdots & f_{n+1m} \end{pmatrix}$$

矩阵距离为

$$d_l=(d_1,d_2,...,d_m)$$

表示第 l 个用户与新用户的属性比较值。

5.3.3 模型建立和分析求解

本题中由于是对新用户的偏好进行分析，故用 Q 型聚类[9]对新老用户属性特征的相似度进行测量。要用数量化的方法对事物进行分类，就必须用数量化的方法描述事物之间的相似程度。一个事物常常需要用多个变量来刻画。如果对于一群有待分类的样本点需用 p 个变量描述，则每个样本点可以看成是 R^p 空间中的一点。因此，可以用距离来量度样本点间的相似度[10]。

记 Ω 是样本点集，距离 $d(\bullet,\bullet)$ 是 $\Omega\times\Omega\to R^+$ 的一个函数，满足条件：

（1） $d(x,y)\geq 0, x,y\in\Omega$；

（2） $d(x,y)=0$，当且仅当 $x=y$；

（3） $d(x,y)=d(y,x), x,y\in\Omega$；

（4） $d(x,y)\leq d(x,z)+d(z,y), x,y,z\in\Omega$。

由于样本间距离的定义满足正定型、对称性和三角不等式。在聚类分析中，为了尽可能避免变量的多重相关性，可以运用马氏距离来度量新老用户之间属性值的差异。即

$$d(x,y)=\sqrt{(x-y)^T\sum^{-1}(x} \qquad (10)$$

其中，x, y为来自p维总体Z的样本观测值；$\sum$为Z的协方差矩阵，实际中$\sum$往往是未知的，常常需要用样本协方差来估计。马氏距离对一切线性变换是不变的，故不受量纲的影响。

根据以上所述的聚类分析的方法，可以采用加权和方法来比较各老用户和新用户间的相似度，即

$$d_l = sim^0_{ua,ub} = \sum_{i=1}^{m} F(f_{ai}, f_{bi}) * w_i, \qquad (11)$$

其中，f_{ai}、f_{bi}分别表示用户a和b的第i个统计属性值，F是一个属性比较函数，w_i为第i个统计属性在相似性计算中的权重，按照各个属性的重要性来设定，m为人口统计信息包含的属性数量。

计算时，a用户不变，一直是新用户，而b用户是对所有老用户的遍历，算出每位老用户与新用户的相对属性值后，再作比较。比较出与新用户距离最相近的老用户，再根据第一问和第二问中用到的方法，分析出这些老用户最感兴趣的影片，最后将这些影片推荐给新用户。

六、模型评价与推广

6.1 模型一的评价及推广

问题一建立了基于 TOPSIS 法的用户喜好评价模型，此模型有很大的优点和独到之处，体现在对数据的统计处理和模型方法的选择。数据处理上，没有简单地对评分取平均值，而是依据电影类别总数对评分进行倒数加权，这样使评分更加合理。随后加权求和取平均值剔除分数评价中次数因素的影响，降低了两个指标的相关性。方法比较了线性加权法和 TOPSIS 法的特点，并对两个指标进行相关性分析，发现两指标有一定的关联性，故 TOPSIS 法显然优于线性加权法。即我们通过比较差异选择建模方法，是比较合理的，而求解结果得到了十分合理的解释，故认为模型相对来说是比较成功的。

6.2 模型二的评价及推广

模型二的优点：该模型首先依据职业和年龄将用户样本较精确地划分为不同类型用户群体，可靠性增强，充分利用已有信息；在现实中，有些用户会有一些潜在喜欢的不同类型电影，基于用户原有观看数据分析，无法得出全面结论，因此，我们分析了与用户相似人群的喜好，对其所偏好的电影进行推荐，所推荐电影从用户主观偏好电影和同类用户偏好的电影中加权分别取得；在主观和客观推荐分析中，由于同类数据量少，为了使所推荐的电影贴近用户喜好，在模型中，用户主观偏好的权数取得较大。

模型二的缺点：模型二预先假定同职业和年龄接近的用户具有相似的喜好，这很大程度上依赖于数据的准确性和兴趣与职业、年龄的相关性。利用这两个条件对用户人群进行划分可能会丢失一些有用的信息；同时，模型二在划分相似性人群的时候得到的样本数可能很少，此时认为地扩大年龄段会使原来年龄相差较大的几类用户先聚成一类，使得后来得到的相关性分析结果有失偏颇。

模型二的推广：基于协同过滤的用户电影推荐模型，可以实现对大量用户的喜好分析和影片推荐。在获得更多用户信息后可以更好地选定用户分类的主要决定因素，从而得到与用户更相近的邻居用户群体；可以考虑使用用户间的聚类或电影的聚类，减少数据的维度，参考标准有对电影的选择和注册信息，或者利用已知数据对每个用户的未评价电影进行预测打分，补充评分矩阵舍得结果利于处理。

6.3 模型三的评价及推广

问题三建立了基于联合聚类的用户电影推荐系统。在模型二中，模型存在的不足之处是：对用户的职业、年龄只进行了定性分析，聚类，没有量化参与匹配。在新的系统中，我们收集到了较多的用户注册信息，这些信息的属性大多数不是数值型数据，无法直接进行相似度衡量，因此要对各个属性做相应的处理：将“年龄”属性做离散化处理，将“学历”“职业”、“性别”等属性符号化，即可定量的描述出不同类型的相似度指标。该推荐系统的功能还不是很完善，稍作改进，可以推广到音乐、视频、商品、交友等方面，根据用户的资料、输入的关键字或历史浏览记录来分析用户的喜好，推荐出用户最可能感兴趣并接受的建议。

七、参考文献

[1]司守奎，孙玺菁.数学建模算法与应用.北京：国防工业出版社，2013 年，345-350

[2]马宏伟，张光卫，李鹏.协同过滤算法综述.小型微型计算机系统，2009 年，第 30 卷（第 7 期）：1283-1284

[3],孔繁胜.基于多数据源和联合聚类的智能推荐.模式识别与人工智能,2008 年，第 21 卷（第 6 期）：776-778

[4]丛明珠，王富喜.城市物流业核心竞争力评价方法及应用.技术与方法，2010 年，第 218 期

[5]周亚，多属性决策中 TOPSIS 的研究.武汉：武汉理工大学出版社，2009 年，6-14

[6]Bhatt Chidansh Amitkumar, Kankanhalli Mohan S. Multimedia Data Ming stare of art and challenges. Multimedia Tools and Applications, 2011, Vol(51).1:35-76

[7]谢发川，基于数据挖掘的视频推荐系统建模研究.成都：电子科技大学，2012 年，7-20

[8]岳文君，一种智能推荐系统的研究与应用，北京；北京邮电大学出版社，2013

[9] Robert K. Merton. The Matthew Effect in Science. Science, 1968, 159(3810):56-63

Rakesh Agrawal, Ramakrishnan Srikant. Fast Algorithms for Mining Association Rules in large database. In Proc. Of the 20th International Conference on Very Large Databases, 1994, 487-499
[10]冷亚军，梁昌勇，丁勇，陆青.协同过滤中一种有效的最近邻选择方法.模式识别与人工智能，2013 年，第 26 卷（第 10 期）；696-673