基于项目综合相似度的协同过滤算法

许智宏，王宝莹
（河北工业大学 计算机与软件学院，天津 300401）

摘　要：提出了一种基于项目综合相似度的协同过滤算法。综合相似度是项目相似度和类别相似度进行加权，加权方式是从热能学中协同计算燃烧传热量的高温辐射换热综合发射率 ε 公式比拟得出，两者均是计算综合系数，在计算综合系数中可以通用。实验结果表明，在推荐不同的前 N 个项目的实验中，用新方法得到的准确率高于传统方法；在固定推荐数目改变最近邻的实验中，用新方法得到的准确率高于传统方法，因此可以得出结论：基于项目综合相似度的协同过滤算法可以提高计算准确性，提高推荐质量。

关键词：协同过滤；项目相似度；类别相似度；综合相似度；发射率
中图分类号：TP391.3；TP301.6　　**文献标志码**：A　　**文章编号**：1001-3695(2014)02-0398-03
doi:10.3969/j.issn.1001-3695.2014.02.018

Collaborative filtering algorithm based on item complex similarity

XU Zhi-hong , WANG Bao-ying
(*School of Computer Science & Engineering*, *Hebei University of Technology*, *Tianjin* 300401, *China*)

Abstract: Proposing a new collaborative filtering algorithm based complex similarity of the item. The similarity is an item similarity and category similarity weighting. It derived the algorithm from the energy science emissivity ε formula which came from the collaborative computing combustion heat of the energy science. Both of them were calculated complex coefficient which could be common in the calculation of the coefficient. Experimental results show that the accuracy obtains with the new method is higher than traditional methods in the experiments of recommending N items. A higher accuracy obtains by using a new method in the fix recommendation and change numbers of the nearest neighbors in the experiment. Therefore it can be concluded that using the complex similarity based the item can improve the accuracy and improve the quality of recommendation.

Key words: collaborative filtering; item similarity; category similarity; complex similarity; emissivity

互联网络时代用户需求日趋多样化、个性化，海量资源的搜索成为研究用户需求的重点问题。对于明确目标关键词的搜索技术已日趋成熟，针对用户个性化需求的推荐系统仍在进行。常见的推荐算法有基于内容的推荐、协同过滤推荐、基于知识的推荐、基于网络结构的推荐及组合推荐等[1]。基于内容的推荐是根据用户已经选择的对象，从推荐对象中选择其他特征相似的对象作为推荐结果，文本信息的特征更易于提取，因而基于内容的推荐更适用于新闻网站，不适用于多媒体推荐，对于多媒体的推荐应用更多的是协同过滤算法[2]。

协同过滤算法包括基于内容的协同过滤和基于项目的协同过滤。协同过滤算法分析用户兴趣，在用户群中找到指定用户的相似兴趣用户，综合这些相似用户对某一信息的评价，形成系统对该指定用户对此信息的喜好程度预测。其中基于项目的协同过滤在图书、电子商务和电影网站的推荐上具有明显的优势，该算法主要在于维护一张项目相关度的矩阵，根据网站每天对于该项目的更新，对相似度矩阵进行更新，以实现对用户的个性化推荐。

针对传统的基于项目的协同过滤算法中存在的推荐质量问题，提出了一种由热能学中发射率 ε 公式类比出的基于类别相似度的相似性协同过滤计算方法，以得到更高的推荐质量。

1　相似性度量方法

1.1　传统的相似性度量方法

基于项目的协同过滤算法的推荐首先是找出当前用户的未评分项目的相似项目。传统的计算相似性方法有余弦相似性、相关相似性和改进的余弦相似性。

1.1.1　余弦相似性

把向量之间的夹角余弦值作为两个项目的相似程度。r_{ui} 代表用户 u 对项目 i 的评分，r_{uj} 代表用户 u 对项目 j 的评分。项目 i 与 j 之间的相似度 $\mathrm{sim}(i,j)$ 表达式[3]为

$$\mathrm{sim}(i,j)=\cos(i,j)=\frac{i\cdot j}{\|i\|\times\|j\|}=\frac{\sum_{u\in U}r_{ui}\cdot\sum_{u\in U}r_{uj}}{\sqrt{\sum_{u\in U}r_{ui}^2}\sqrt{\sum_{u\in U}r_{uj}^2}} \tag{1}$$

1.1.2　相关相似性

皮尔逊相关系数一般用于计算两个定距变量间联系的紧密程度，由于其容易理解便于计算而被广泛使用。其具体表达式[4]为

$$\mathrm{sim}(i,j)=\frac{\sum_{u\in U}(r_{ui}-r_i)\cdot\sum_{u\in U}(r_{uj}-r_j)}{\sqrt{\sum_{u\in U}(r_{ui}-r_i)^2}\sqrt{\sum_{u\in U}(r_{uj}-r_j)^2}} \tag{2}$$

收稿日期：2013-05-22；**修回日期**：2013-06-24

作者简介：许智宏(1970-)，女，天津人，副教授，博士，主要研究方向为分布式技术、智能信息处理(xuzhihong@scse.hebut.edu.cn)；王宝莹(1989-)，女，硕士，主要研究方向为网格与协同计算.

1.1.3　修正的余弦相似性

不同用户的评分尺度是不一样的,显然式(1)中并没有考虑到。评分尺度是指不同用户的评分水平有差别,如有些用户喜欢给高分,而有些用户比较苛求只给低分。为了避免这种情况,一种修正的余弦相似性被提出,其具体如式(3)所示[5]。

$$\mathrm{sim}(i,j)=\frac{\sum_{u\in U}(r_{ui}-r_u)\cdot\sum_{u\in U}(r_{uj}-r_u)}{\sqrt{\sum_{u\in U}(r_{ui}-r_u)^2}\sqrt{\sum_{u\in U}(r_{uj}-r_u)^2}} \tag{3}$$

1.2　基于类别相似度的相似性计算

采用 Movielens 站点提供的数据集,该数据集包括 943 个用户对 1 682 部电影 100 000 条评分,评分取值为(1,2,3,4,5)中任一个数,值越大说明用户喜好程度越高,每个用户至少评价了 20 部电影,电影已有类别 19 类。在原始数据中随机抽取每一位用户的 10 条评论形成 9 430 条评论的测试集,余下的形成训练集。其中 base 为训练集,而 test 为测试集。每一组 . base与. test 文件相对应,为其中一组实验。

从传统的基于项目的协同过滤算法的相似性计算过程中可以发现,虽然其充分地利用了高维稀疏的用户—项目评分数据矩阵,但是对于项目的类别属性却没有关注过。利用项目的类别进行推荐计算,数据进行分类处理也就是通过已有的原始数据中电影项目的 19 个分类来计算项目之间的相似性。项目 i 与 j 之间的类别相似度用 $\mathrm{sim}_p(i,j)$ 来表示,基于用户—项目评分矩阵的项目相似度用 $\mathrm{sim}_r(i,j)$ 来表示。综合相似度考虑了这两种相似性之间的关系。综合相似度的计算形式是由航天发动机推力室协同计算燃烧传热量,考虑到高温辐射换热计算时所有壁面的综合发射率 ε 公式[6] 联想到的,两者均是计算综合系数,相似度取值均在(0,1)之间。发射率公式计算面壁间的距离,综合相似度计算类别间的距离。发射率 ε 公式形式如式(4)所示。

$$\varepsilon=\frac{\varepsilon_i}{1+\varepsilon_i\left(\frac{1}{\varepsilon_j}-1\right)} \tag{4}$$

其中:ε_i 为壁面 i 的发射率,ε_j 为壁面 j 的发射率。ε_i 与 ε_j 均为温度 T 与辐射波长 λ 的二元函数,其范围在(0,1),其中$\frac{\varepsilon_i}{\varepsilon_j}-\varepsilon_i$ 表示壁面 i 与 j 发射率的距离性或者这两个壁面之间辐射热阻比,壁面的发射率 ε_i 可以与项目相似性 $\mathrm{sim}_r(i,j)$ 类比,壁面 j 的发射率 ε_j 可以与类别相似性 $\mathrm{sim}_p(i,j)$ 类比,由此联想出式(5)(6)。

$$\mathrm{sim}(i,j)=\frac{\mathrm{sim}_r(i,j)}{1+c\ \mathrm{sim}_p(i,j)} \tag{5}$$

$$c\ \mathrm{sim}_p(i,j)=1-\mathrm{sim}_p(i,j) \tag{6}$$

其中:$c\ \mathrm{sim}_p(i,j)$ 为 $\mathrm{sim}_p(i,j)$ 的余补,表示项目 i 与 j 的类别不相关性。考虑到项目类别对项目之间相似性的影响,所以将基于用户—项目评分矩阵的评分相似度 $\mathrm{sim}_r(i,j)$ 用类别不相关性进行修正。如果项目 i 与 j 的不相关性 $c\ \mathrm{sim}_p(i,j)$ 为 0,也就是说它们的类别完全相同,则此时基于项目相似度 $\mathrm{sim}_r(i,j)$ 是在同一个类别集合中进行计算。若项目 i 与 j 分别属于两个不同的类别,此时不相关性不为 0,就视这两个类别的交集情况进行修正。若不相关性 $c\ \mathrm{sim}_p(i,j)$ 为 1 也就是项目 i 与 j 的类别相似度 $\mathrm{sim}_p(i,j)$ 为 0,此时可以肯定项目 i 与 j 所在的类别没有交集,这样它们之间的项目相似度是建立在两个完全不同的类别中进行的,所以用 $\mathrm{sim}_r(i,j)$ 去除以 2。

建立项目—类别矩阵 $\boldsymbol{Q}_{n\times p}$,其中 n 为项目数量,p 为类别数量。将项目的类别数据视为 p 维空间上的向量,这样计算两个项目之间的类别相似度就是计算这两个向量之间的夹角余弦值,余弦相似度即式(1),具体计算公式[7] 如式(7)所示。

$$\mathrm{sim}(i,j)=\cos(i,j)=\frac{i\cdot j}{\|i\|\times\|j\|}=\frac{\sum_{t\in p}\lambda_{it}\cdot\lambda_{jt}}{\sqrt{\sum_{t\in p}\lambda_{it}^2}\times\sqrt{\sum_{t\in p}\lambda_{jt}^2}} \tag{7}$$

其中:P 是项目总数,t 代表具体某个类别。在本文中 $P=19$,而 t 的取值分别是 0 ~ 18。λ_{it} 与 λ_{jt} 分别表示项目 i 与 j 对第 t 个类别的归属情况,归属为 1,不归属为 0,具体是对 19 个类别的归属情况。

2　基于项目类别相似度的协同过滤算法

利用用户项目矩阵和式(2)得到项目之间的相似性,利用项目类别矩阵和式(7)得到类别之间的相似性,利用综合相似度式(5)(6)得到项目相似矩阵,比较相似度大小,得到降序排序的数据,即项目最近邻。通过最近邻预测未评分项目的评分,选取 Top-N 即可进行推荐。具体过程如图 1 所示。

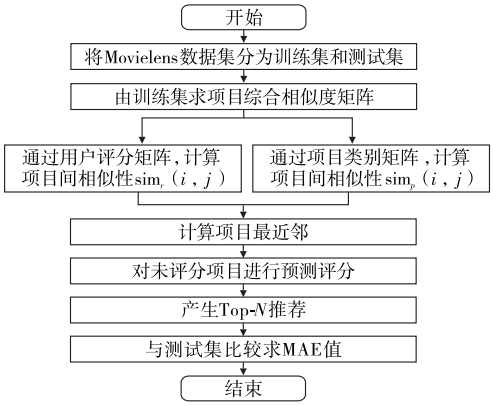

图 1　综合相似度推荐过程

2.1　项目最近邻

求项目最近邻首先要求项目相似度。相似性计算方法是综合了项目相似性和类别相似性一起计算的式(5),其中项目间的相似性用式(2)计算,项目所属类别间的相似性用式(7)计算。求项目相似度以及最近邻伪代码如下:

输入:用户评分矩阵 $\boldsymbol{R}_{mn}$,项目类别矩阵 $\boldsymbol{R}_{np}$,推荐集元素 N。

输出:最近邻矩阵 Neighbour。

for 项目 i、j,类别 k,do

　利用式(2)计算项目 i、j 间的相似性 $\mathrm{sim}_r(i,j)$

　利用式(5)计算类别间相似性 $\mathrm{sim}_p(i,j)$

　综合以上两种相似性度量方法,利用式(4)计算 $\mathrm{sim}(i,j)$

　更新用户的相似度矩阵 sim

for 每个项目 i do

寻找最近邻项目集合 Neighbour = $\{j_1,j_2,\cdots,j_k\}$,且满足 $\mathrm{sim}(i,j_1)\geqslant\mathrm{sim}(i,j_2)$。

依次将每一个项目视为目标项目,i 为 1 ~ 1682,分别计算出该项目与其他项目的综合相似度,从而形成 1682 × 1682 的综合评分相似度矩阵 sim。然后对于矩阵 sim 每一行中的元素进行比较,在计算出目标项目 i 与其他项目之间的综合相似度后进行比较,选出其中最大综合相似度的 M 个项目作为 i 的最邻近集合,并记住这 M 个项目的编号,将相应的编号放入矩阵 $\boldsymbol{g}$ 中,其中 neighbor 与 $\boldsymbol{g}$ 为 $1682\times M$ 矩阵,记住项目编号。

2.2 产生推荐

I 为目标项目 i 的最邻近项目集合，且 $I=\{i_1,i_2,i_3\cdots i_t\}$，其中 $i\notin I$，并有 $\mathrm{sim}(i,i_{t-1})\geqslant\mathrm{sim}(i,i_t)$，从而目标用户的预测评分[8]可以表示为

$$P=\frac{\sum_{i_t\in I}\mathrm{sim}(i,i_t)P_{ui_t}}{\sum_{i_t\in I}|\mathrm{sim}(i,i_t)|} \tag{8}$$

其中：P_{ui} 为用户 u 对目标项目 i 的预测评分。通过预测评分计算出目标用户 u 对其所有未评分项目的预测评分 P_{ui}，并将 P_{ui} 数值从大到小进行排列，从中选出前 N 个最大的评分。

3 实验结果分析

3.1 度量标准

评分预测：设 P_i 为目标用户 u 对测试集中项目 i 的预测评分，而 q_i 为测试集中用户 u 对项目 i 的真实评分。平均绝对误差(mean absolute difference，MAE)是通过将目标用户预测评分与目标用户实际评分进行对比，以两者之间的偏差度来表示推荐效果的准确性。MAE 值越低，说明推荐的方法越准确。如果目标用户在测试集中共有 k 条评分数据，记为 $\{q_1,q_2,\cdots,q_k\}$，目标用户相应的预测评分数据记为 $\{P_1,P_2,\cdots,P_k\}$，则 MAE 计算公式[9]可以表示为

$$\mathrm{MAE}=\frac{\sum_k|p_k-q_k|}{k} \tag{9}$$

3.2 实验结果

实验中的变量有 top-N 和最近邻[10]，第一组实验中最近邻值是固定的 1 000，选取 top-N 不同变量时，会得到不同的 MAE 值，它们一一对应，并且有一定规律。本文通过实验观察不同 N 时的 MAE 值变化。第二组实验是当得到效果较好的 N 时，固定 N 值，选取不同的最近邻值得到不同的 MAE 值，观察规律，得出不同近邻与 MAE 值之间的关系。

3.2.1 top-N 变化时的结果

将推荐结果的 top-N 分别设定为 20、25、30、35、40、45、50。在项目最邻近数量设定为 1 000 个条件下进行实验计算，并且得出传统算法与改进算法的 MAE 值之间直方图的比较，其具体数值如表 1 所示。考虑了类别相似度的改进算法的 MAE 值平均提高了 6.45%。

表 1 top-N 的具体 MAE 值

推荐数量	传统算法的 MAE	改进后算法的 MAE	精度提高百分比
20	0.838 9	0.787 4	6.14
25	0.824 8	0.799 3	3.09
30	0.858 1	0.776 9	9.46
35	0.820 7	0.784 8	8.92
40	0.822 4	0.749 0	8.93
45	0.802 6	0.772 4	3.76
50	0.814 9	0.775 5	4.83

3.2.2 最邻近变化时的结果

在实验中固定 top-N 为 30，将最邻近集合元素个数分别设定为 500、600、700、800、900、1 000 进行实验。运行出传统算法与本文改进算法的 MAE 值，同时生成直方图进行对比。直方图如图 2 所示，具体值如表 2 所示。

传统算法的 MAE 值大多处于 0.8 之上，只有在最近邻数为 600 与 700 时处于 0.8 之下，但也要比同条件下改进的 MAE 值大，精度平均提高了 4.46%。

表 2 不同最近邻的具体 MAE 值

推荐数量	传统算法的 MAE	改进后算法的 MAE	精度提高百分比
1 000	0.858 1	0.776 9	9.46
900	0.809 4	0.794 4	1.85
800	0.807 0	0.787 1	2.47
700	0.790 2	0.767 3	2.90
600	0.791 9	0.782 8	1.15
500	0.858 5	0.781 6	8.96

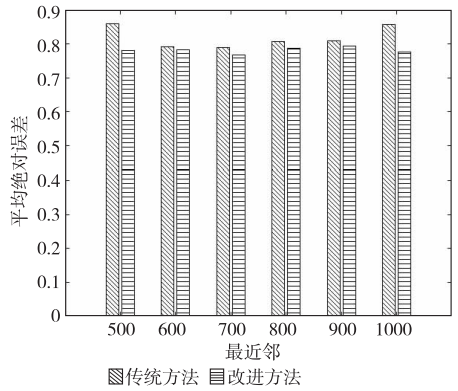

图 2 不同最近邻的 MAE 值

4 结束语

传统的基于项目的协同过滤推荐算法未能充分利用项目在类别上的相似性，因此在项目相似性度量上不够准确。综合相似度将项目类别相似性与传统的项目评分相似性进行加权组合，在此基础上寻找目标项目的最近邻居项目集合，提高了最近邻居项目搜寻的准确度。实验结果表明，新算法能有效提高推荐质量，可以在推荐网站中被推广。

参考文献：

[1] 王国霞，刘贺平. 个性化推荐系统综述[J]. 计算机工程与应用，2012，48(7)：66-76.

[2] BARALIS E，GARZA P. Item selection for associative classification [J]. International Journal of Intelligent Systems，2012，27(3)：279-299.

[3] YE Jun. Cosine similarity measures for intuitionistic fuzzy sets and their applications [J]. Mathematical and Computer Modelling，2011，53(1-2)：91-97.

[4] GONG Song-jie，YE Hong-wu. Joining user clustering and item based collaborative filtering in personalized recommendation services[C]// Proc of International Conference on Industrial and Information Systems. Washington DC：IEEE Computer Society，2009：149-151.

[5] 罗辛，欧阳元新，熊璋，等. 通过相似度支持度优化基于 K 近邻的协同过滤算法[J]. 计算机学报，2010，33(8)：1437-1445.

[6] 杨世铭，陶文铨. 传热学 [M]. 西安：西北工业大学出版社，2012：28-36.

[7] 田丰，桂小林，杨攀，等. 基于类别相似度聚合的关联文本分类方法[J]. 西安交通大学学报，2012，46(12)：6-11.

[8] JIA Rong-fei，JIN Mao-zhong，LIU Chao. A new clustering method for collaborative filtering[C]//Proc of International Conference on Networking and Information Technology. 2010：488-492.

[9] 陈永平，杨思春，刘俞. 基于项目内容和评分的时间加权协作过滤算法[J]. 苏州科技学院学报：自然科学版，2013，3(1)：65-70.

[10] 陶俊，张宁. 基于用户兴趣分类的协同过滤推荐算法[J]. 计算机系统应用，2011，20(5)：55-59.