在 5~40 之间, 进行多组实验.

MovieLens 的用户集中包含了性别、年龄、职业三项用户属性, 本文分析这三项特征信息, 计算出用户之间的特征差$attr(u,v)$如式 (16).

$$attr(u,v)=\alpha\cdot sex+\beta\cdot age+\gamma\cdot occupation \tag{16}$$

本文的 α、β、γ 皆取 1/3, 满足式 (6) 的条件.

对 User.data 的预处理:

(1) 性别判定

MovieLens 数据集中男性用户性别表示为 M, 女性用户性别表示为 F. 若两用户性别相同, *sex* 取值为 0, 若不同, *sex* 取值为 1.

(2) 年龄量化 (表 4)

表 4 MovieLens 用户年龄的量化

年龄	7–17	18–25	26–35	36–45	46–60	61–73
量化值	1	2	3	4	5	6

(3) 职业量化 (表 5)

表 5 MovieLens 用户职业的量化

职业类别	other	Academic/educator	artist	clerical/admin	college/grad student	customer service	doctor/health care
量化值	0	1	2	3	4	5	6
职业类别	executive/managerial	farmer	homemaker	K-12 student	lawyer	programmer	retired
量化值	7	8	9	10	11	12	13
职业类别	sales/marketing	scientist	self-employed	technician/engineer	tradesman/craftsman	unemployed	writer
量化值	14	15	16	17	18	19	20

由此, 式 (14) 中 *sex* 取值为 0 或 1, *age*, *occupation* 的取值为两用户特征值量化后差值的绝对值.

求得用户特征差值$attr(u,v)$后, 利用 Sigmoid 函数, 计算用户 u 和用户 v 之间的用户特征信息相似度 $sim_{attr}(u,v)$, 如式 (8).

传统协同过滤算法采用改进的余弦相似度 User-IIF 算法[15], 如式 (17) 所示, 降低了用户 u 和用户 v 共同兴趣列表中热门物品对他们相似度的影响.

$$w_{uv}=\frac{\sum_{i\in N_{(u)}\cap N_{(v)}}\dfrac{1}{\log(1+|N(i)|)}}{\sqrt{|N(u)||N(v)|}} \tag{17}$$

求得相似度后, 计算预测评分, 为用户进行推荐. 进行多次实验, 求得多组评估指标, 并用 Matlab 进行仿真, 仿真结果如图 4、图 5 所示.

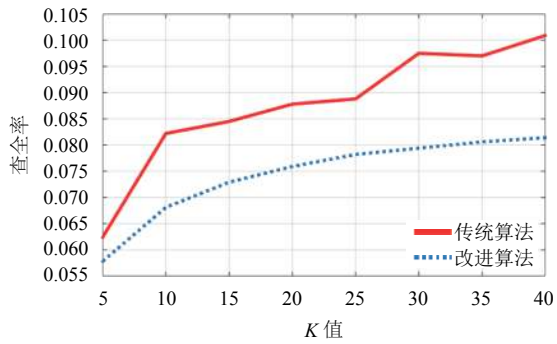

图 4 改进前后算法查全率随 K 值的变化

图 4 表示基于用户的协同过滤算法在新用户的情况下, 新算法与传统算法的查全率与邻居数的变化关系图. 由图可知, 随着邻居数 K 值增加, 系统的查全率呈上升趋势, 并且新算法的查全率高于传统算法, 说明新算法的检索结果更有效.

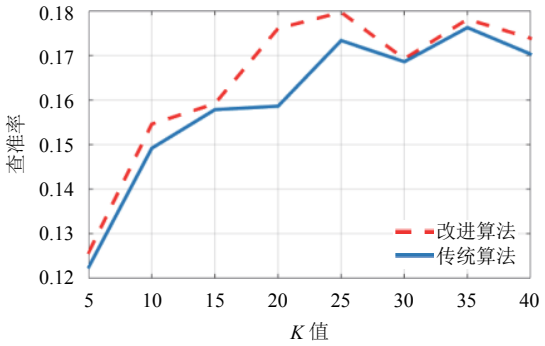

图 5 改进前后算法查准率随 K 值的变化

图 5 表示基于用户的协同过滤算法, 在新用户的情况下新算法与传统算法查准率与 K 值的关系. 可以看出, 新算法推荐结果的准确率高于传统算法. 改进后算法推荐精度更高, 有效地改善了系统新用户的冷启动问题.

结合图 4、图 5 发现, 在改进后的 User-based CF 中, 当邻居数取 35 时算法查全率最大, 查准率也比较高, 算法推荐质量较好.

3.3.2 新项目冷启动算法验证

本文提出的新项目的冷启动算法采用了凝聚式层

次聚类的思想, 在 MovieLens 数据集上进行实验结果验证, 将训练集和测试集的比例分为 9:1, 新项目和老项目的比例为 3:7.

Step 1. 对 movie.data 的数据初始化处理.

(1) 年份 (year) 关键词: 直接表示为 i_y、j_y.

(2) 派别 (genres) 关键词: 遍历电影所属的派别, 若两部电影有属于一个相同的派别, 则 g 值减 1, 否则, g 值保持不变 (g 初始值为 3), 最后得到目标电影和其他电影派别上的距离 g 值.

Step 2. 欧式距离的计算.

在对电影的基本数据处理之后, 计算电影之间的欧几里德距离. 欧式距离计算如式 (10).

Step 3. 层次聚类.

最后选取不同的邻居数 K 值在 5~40 之间, 进行多次实验比较加入新项目时, 改进前后基于项目的协同过滤算法在推荐精度上的变化, 如图 6、图 7 所示.

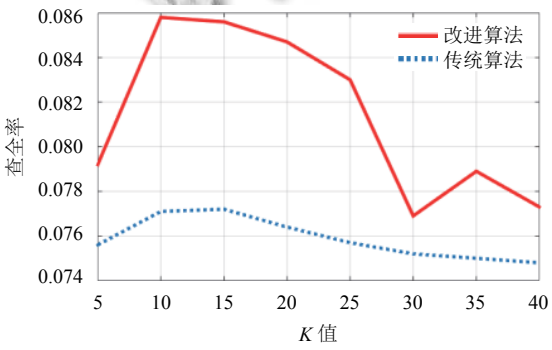

图 6 改进前后 ItemCF 算法查全率随 K 值的变化

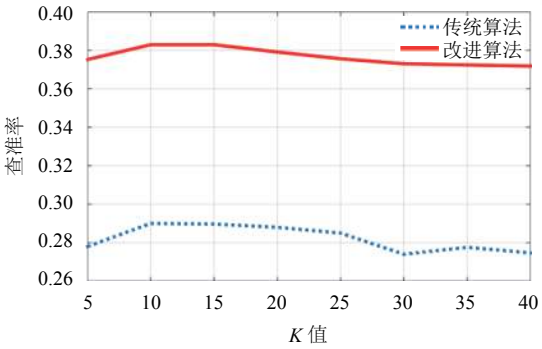

图 7 改进前后 ItemCF 算法查准率随 K 值的变化

传统方法采用的相似度计算公式如式 (18), 减轻了热门物品和其他众多物品相似的可能性.

$$w_{ij}=\frac{|N(i)\cap N(j)|}{\sqrt{|N(i)||N(j)|}} \tag{18}$$

从图 6 可以看出, 采用了层次聚类的算法在查全率上优于传统算法, 在邻居数取 10 的时候, 查全率达到最大值, 在邻居数取 25~35 之间时, 查全率波动较大, 并且呈下降趋势. 整体上看, 改进后的层次聚类算法推荐结果中被检索到的更多, 查全率更高.

图 7 比较了采用凝聚式层次聚类的协同过滤和传统基于项目的协同过滤算法在查准率上的性能, 由图可见, 算法的查准率比较平稳, 在加入了新项目后, 改进后算法的查准率优于传统算法的值, 表现出更好的推荐精度.

结合图 6、图 7, 在改进后的 Item-based CF 中, 邻居数取 10 时, 算法的推荐性能较好.

4 结论

本文首先介绍了协同过滤算法以及算法的冷启动问题, 重点对新用户和新项目的冷启动问题进行研究, 提出了融合用户信息模型的基于用户的协同过滤算法和采用层次聚类的基于项目的协同过滤算法. 具体研究工作如下:

(1) 针对新用户的冷启动, 算法提取了用户个人特征信息, 为用户信息建模, 调用 Sigmoid 函数求得基于用户特征模型的相似度.

(2) 对于新项目的冷启动, 算法提取项目的信息属性, 计算出欧式距离, 采用凝聚式层次聚类的方法, 找到目标项目的邻居项目, 计算预测评分, 完成推荐.

(3) 选用网络开源数据集 MovieLens 进行实验验证, 将新算法与传统算法多次实验对比. 结果表明, 在新用户和新项目的情况下, 新算法推荐结果的查全率、查准率都有所提升, 有效地缓解了传统协同过滤算法的冷启动问题, 改善了推荐质量.

参考文献

1 Maes P. Agents that reduce work and information overload. Communications of the ACM, 1994, 37(7): 30–40. [doi: 10.1145/176789.176792]

2 Guo YY, Liu QC. E-commerce personalized recommendation system based on multi-agent. Proceedings of the 7th International Conference on Fuzzy Systems and Knowledge Discovery. Yantai, China. 2010. 1999–2003.

3 Bobadilla J, Ortega F, Hernando A. A collaborative filtering similarity measure based on singularities. Information Processing & Management, 2012, 48(2): 204–217.

4 Hu R, Pu P. Enhancing collaborative filtering systems with personality information. Proceedings of the 5th ACM Conference on Recommender Systems. Chicago, IL, USA. 2011. 197–204.
5 Wang JW. A collaborative filtering systems based on personality information. Proceedings of 2015 International Industrial Informatics and Computer Engineering Conference. Xi'an, China. 2015. [doi: 10.2991/iiicec-15.2015.163]
6 Goldberg D, Nichols D, Oki BM, *et al*. Using collaborative filtering to weave an information tapestry. Communications of the ACM, 1992, 35(12): 61–70. [doi: 10.1145/138859.138867]
7 申辉繁. 协同过滤算法中冷启动问题的研究[硕士学位论文]. 重庆: 重庆大学, 2015.
8 马卓. 协同过滤推荐算法的研究与改进[硕士学位论文]. 秦皇岛: 燕山大学, 2015.
9 Xu JW, Yao Y, Tong HH, *et al*. RaPare: A generic strategy for cold-start rating prediction problem. IEEE Transactions on Knowledge and Data Engineering, 2017, 29(6): 1296–1309. [doi: 10.1109/TKDE.2016.2615039]
10 Nguyen VD, Sriboonchitta S, Huynh VN. Using community preference for overcoming sparsity and cold-start problems in collaborative filtering system offering soft ratings. Electronic Commerce Research and Applications, 2017, 26: 101–108. [doi: 10.1016/j.elerap.2017.10.002]
11 Katarya R, Verma OP. Effective collaborative movie recommender system using asymmetric user similarity and matrix factorization. Proceedings of 2016 International Conference on Computing, Communication and Automation. Noida, India. 2016. 71–75.
12 乔雨, 李玲娟. 推荐系统冷启动问题解决策略研究. 计算机技术与发展, 2018, 28(2): 83–87. [doi: 10.3969/j.issn.1673-629X.2018.02.019]
13 http://www.grouplens.org/node/73.
14 程淑玉. 基于协同过滤算法的个性化推荐系统的研究[硕士学位论文]. 合肥: 合肥工业大学, 2010.
15 Breese JS, Heckerman D, Kadie C. Empirical analysis of predictive algorithms for collaborative filtering. Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence. Madison, WI, USA. 1998. 43–52.