

基于用户特征和项目属性的协同过滤推荐算法

陈志敏, 李志强

(扬州大学 信息工程学院, 江苏 扬州 225009)

(zmchen@yzu.edu.cn)

摘要:在数据极度稀疏的环境下, 仅仅依赖用户直接评分数据的传统协同过滤算法无法取得满意的推荐质量。提出基于用户特征和项目属性的协同过滤算法, 在用户相似性计算过程中引入时间相关的兴趣度, 使得最近邻的确定更加准确; 预测评分时, 通过衡量用户信任度来体现各邻居对目标用户最终推荐的贡献程度, 并以用户对项目属性的偏好度代替评分数据对新项目进行推荐。基于 MovieLens 数据集进行的实验结果表明, 改进后的算法有效解决了系统冷启动问题, 明显提高了系统推荐的准确度。

关键词:协同过滤; 相似性计算; 用户特征; 冷启动

中图分类号:TP311 **文献标志码:**A

Collaborative filtering recommendation algorithm based on user characteristics and item attributes

CHEN Zhi-min, LI Zhi-qiang

(College of Information Engineering, Yangzhou University, Yangzhou Jiangsu 225009, China)

Abstract: Under the extremely sparse data environment, the traditional collaborative filtering algorithms only depending on users rating data cannot achieve satisfactory recommended quality. A recommendation algorithm based on user characteristics and item attributes was provided. First, the time-related interest degree was introduced in the process of user similarity calculation, which made a more accurate nearest neighbor set. While predicting the rating for the target user, the trust measure was used to reflect the neighbors' contribution level for the ultimate recommendation. In addition, the users' preference on item attribute instead of rating score was used to recommend the new items. The experimental results based on MovieLens data set show that the improved algorithm can solve the problem of cold-start and improve the accuracy of system recommendation significantly.

Key words: Collaborative Filtering (CF); similarity measure; user characteristic; cold-start

0 引言

为解决当前互联网上信息过载问题, 很多大型电子商务网站都不同程度地使用了各种形式的推荐系统^[1-2]。协同过滤(Collaborative Filtering, CF)被认为是迄今为止最成功、应用最广泛的个性化推荐技术之一, 它的核心思想是基于用户对项目的评分度量用户间的相似性, 寻找与目标用户兴趣最相似的邻居用户, 然后利用邻居用户对项目的评分来预测目标用户对该项目的喜好程度, 从而产生推荐^[3-4]。

传统的协同过滤推荐算法在度量用户相似性时, 只注重用户之间评分数据的相似性, 并未考虑用户兴趣是随时间推移而发生变化的, 将用户在不同时期的评分等同对待, 导致推荐的信息偏离用户的当前需求。此外, 评分数据的真实可靠也是影响推荐质量的一个重要因素, 不真实的评分数据会降低预测的准确性和推荐性能。而且随着系统规模的进一步扩大, 项目空间上的用户评分数据变得极端稀疏, 尤其当一个新项目进入系统后, 由于还未接受任何用户的评分而导致无法作出推荐, 即冷启动问题^[5], 严重影响了系统的推荐质量。针对上述问题, 在传统算法的基础上, 本文提出了基于用户特征和项目属性的协同过滤算法, 通过引入用户兴趣度来体现用户需求随时间变化的实际情况, 通过度量邻居用户评分的权威

度、公正度和准确度来衡量他们相对于目标用户的推荐信任度, 并利用用户对项目属性的偏好度来实现新项目的推荐。

1 传统协同过滤算法

在传统基于用户协同过滤推荐算法的系统中, 数据核心是一个用户-项目评分矩阵 $R_{m \times n}$, 其中 m 表示用户数, n 表示项目数, $R_{i,j}$ 表示第 i 个用户对第 j 个项目的评分, 即喜好程度, 通常取值为 1 ~ 5, 若用户未对该项目评分, 取值为 0。

1.1 邻居形成

目标用户的邻居形成是协同过滤算法中的关键步骤, 主要通过衡量用户之间的相似程度, 找出与目标用户具有相似爱好的最近邻居集合。相似程度的度量方法主要有余弦相似性和 Pearson 相关系数^[6]。

余弦相似性 将用户评分看做 n 维项目空间上的向量, 用户间的相似性通过向量间的夹角余弦进行度量。设用户 u 和用户 v 在项目空间上的评分向量分别为 $u = (R_{u,1}, R_{u,2}, \dots, R_{u,n})$ 和 $v = (R_{v,1}, R_{v,2}, \dots, R_{v,n})$, 则两者间的相似性 $\text{sim}(u, v)$ 可按式(1)计算:

$$\text{sim}(u, v) = \cos(u, v) = \frac{u \cdot v}{\|u\| \cdot \|v\|} \quad (1)$$

其中: $u \cdot v$ 表示用户向量 u 和 v 的内积, $\|u\|$ 和 $\|v\|$ 分别

收稿日期:2011-01-21;修回日期:2011-03-07。 基金项目:国家自然科学基金资助项目(61070240)。

作者简介:陈志敏(1976-),女,江苏扬州人,讲师,硕士,主要研究方向:Web数据挖掘;李志强(1974-),男,江苏泰州人,副教授,博士,主要研究方向:算法优化。

表示两者的模。用户向量间的夹角越小,说明两者的相似程度越高,反之相似程度越低。

Pearson 相关系数 在用户共同评分项目的基础上度量用户间的相似度,设 $I(u) \cap I(v)$ 为用户 u 和用户 v 共同评分的项目集合,两用户间的相似性 $\text{sim}(u, v)$ 表示如下:

$$\text{sim}(u, v) = \frac{\sum_{i \in I(u) \cap I(v)} (R_{u,i} - \bar{R}_u)(R_{v,i} - \bar{R}_v)}{\sqrt{\sum_{i \in I(u) \cap I(v)} (R_{u,i} - \bar{R}_u)^2} \sqrt{\sum_{i \in I(u) \cap I(v)} (R_{v,i} - \bar{R}_v)^2}} \quad (2)$$

其中, $R_{u,i}$ 和 $R_{v,i}$ 分别表示用户 u 和用户 v 对项目 i 的评分值, \bar{R}_u 和 \bar{R}_v 分别表示用户 u 和用户 v 的评分均值。大量实验表明, Pearson 相关系数能更好地衡量用户或项目间的相似程度^[7]。

1.2 推荐产生

设目标用户 a 的最近邻居集合为 N_a , 则用户 a 对未评分项目 i 的预测评分 $r_{a,i}$ 可以通过所有邻居用户对项目 i 评分的加权平均值逼近, 计算方法如下所示:

$$r_{a,i} = \bar{r}_a + \frac{\sum_{j \in N_a} (R_{j,i} - \bar{R}_j) \times \text{sim}(a, j)}{\sum_{j \in N_a} |\text{sim}(a, j)|} \quad (3)$$

2 基于用户特征的协同过滤算法

2.1 用户兴趣度

传统基于用户的 CF 算法主要利用兴趣相似的邻居用户对某个项目的评分来预测当前用户对该项目的喜好程度, 而用户的兴趣偏好是会随时间变化的, 在不同时刻对同一项目的评分也不尽相同。而传统算法将用户不同时间的项目评分同等对待, 导致寻找到的邻居用户不够准确, 从而降低了系统的推荐质量。一般来说, 用户感兴趣的项目最可能是他近期访问过的项目, 因此近期评价过的项目对推荐结果的贡献较大, 而早期的评价数据对生成推荐的影响相对较小, 因此在同一或相近时间段内衡量目标用户与其他用户间的评分相似性, 在此基础上确定的最近邻居才是合理准确的^[8]。

为了反映用户兴趣随时间变化的实际情况, 本文引入基于访问时间的数据权重, 以提高最近评分数据在推荐生成过程中的重要性。受遗忘规律的启发, 人在识记后的短时期内遗忘较快, 而经过较长时间后遗忘变得越来越缓慢, 是一个先快后慢的非线性过程。因此本文采用逐步遗忘策略, 按时间对评分的重要性进行不同速度的衰减。假设用户 u 所有评分项目中最早评价时刻为 t_0 , 其中项目 i 的评价时刻为 t_i , 当前时刻为 t_c , 则用户 u 对项目 i 的兴趣度 $I(u, i)$ 可定义如下:

$$I(u, i) = \begin{cases} \frac{e^{-(t_c - t_i)}}{t_c - t_0}, & t_i \neq t_c \\ 1, & t_i = t_c \end{cases} \quad (4)$$

可以发现, 用户兴趣度函数 $I(u, i)$ 是一个逐步递减的非线性函数, 若用户 u 评价项目 i 的时间距当前时间较近, 即 $t_c - t_i$ 值较小时, 所得兴趣度函数的值较大, 表明用户对该项目兴趣较浓, 反之则较弱。因而该兴趣度函数在突出最近评分重要性的同时削弱了过去评分的重要程度, 有效反映了用户的兴趣变化, 能够找出近期兴趣最相似的用户。

2.2 用户信任度

在实际生活中, 当需要向朋友咨询意见时, 会因为对不同朋友信任的差别而影响意见的接受程度^[9]。同样在推荐系统中, 目标用户对邻居用户的信任程度也是影响推荐质量的

一个重要因素。传统推荐算法仅仅注重用户间评分数据的相似性, 推荐有效的前提是参与预测的邻居用户的评分都是真实可靠的, 但实际由于种种原因, 这种前提很难得到保证。比如有些不良商家会通过各种手段伪造恶意评价, 以使自己的产品优先得到推荐。如果这些数据直接被用来参与预测目标用户的评分, 就会严重影响预测值的准确性和推荐系统的性能。为此, 本文引入信任度概念对邻居用户推荐的可信任程度进行评估, 主要从用户评分的权威度、公正度和准确度三方面进行考察。

权威度 考虑到在协同过滤系统中, 有的用户积极主动, 愿意对若干相关项目做出评价, 这些人会因此逐渐赢得其他用户的信任, 所作评分的权威性也相对较高; 而有些用户则消极被动, 不愿或很少为系统提供评价, 这些人的评分对于推荐的贡献就相对较小, 并逐渐失去信任。因此我们可以认为评价数量较多的积极用户评分的权威性要高于评分较少的消极用户。设用户权威度标记为 $A(u)$, 则可定义如下:

$$A(u) = 1 - 1/\ln |N_u| \quad (5)$$

其中: N_u 为用户 u 在系统中所有评分项目的集合, $|N_u|$ 表示集合的大小, 集合中的项目数量越多, 说明该用户的权威度和被信任的程度越高, 反之权威度越低。

公正度 考虑到用户打分时的随意性也会影响评分的客观公正, 因此我们利用项目评分的均方差来衡量用户的评分态度。某用户所有项目评分的均方差越小, 表明这些评分相对稳定, 体现了该用户评分态度的客观公正。若用户 u 的公正度标记为 $E(u)$, 则可定义如下:

$$E(u) = \sqrt{\sum_{i \in N_u} (R_{u,i} - \bar{R}_u)^2 / |N_u|}; |N_u| \neq 0 \quad (6)$$

其中: N_u 意义同式(5), $R_{u,i}$ 表示用户 u 对该集合 N_u 中某个项目 i 的评分, \bar{R}_u 代表用户 u 对集合 N_u 中所有项目评分的平均值。若用户 u 为一个新用户, 即未对系统中任一项目作出评价, 则公正度为 0。

准确度 用以衡量用户对资源项目评分的准确程度, 若某个用户对某个项目的评分与所有用户对该项目评分的平均值越接近, 则该用户评分的准确度就越高, 因而被信任的程度也相应提高。设用户的准确度标记为 $C(u)$, 则计算公式如下:

$$C(u) = 1 / \sum_{i \in N_u} (R_{u,i} - \bar{R}_i); |N_u| \neq 0 \quad (7)$$

其中, N_u 和 $R_{u,i}$ 意义同式(6), \bar{R}_i 代表所有用户对项目 i 评分的平均值。同样当用户 u 为新用户时, 准确度为 0。

在整个项目空间上, 用户评分越权威、公正和准确, 该用户越值得信任, 则用户最终的信任度 $T(u)$ 可以定义为上述三个度量的组合, 表示如下:

$$T(u) = A(u) \times E(u) \times C(u) \quad (8)$$

2.3 用户属性偏好度

随着电子商务规模的扩大, 系统中用户和商品项目的数量也在迅速增加, 每个用户只会对其中的一小部分感兴趣并进行评分, 所评项目只占项目总数的 1% ~ 2%^[10]。评分数据的极度稀疏使得单纯依赖用户评分的传统协同过滤算法的推荐质量严重下降, 特别是当一个新项目进入系统后, 由于还未接受任何用户的评分, 导致该项目无法推荐给对其感兴趣的用户, 即冷启动问题。为此, 本文基于项目自身的属性, 以用户对新项目所具有的相关属性的偏好程度来代替对该新项目的评分, 并以所有邻居用户偏好度的加权平均来预测目标用户对该新项目的评分。一般电子商务系统中商品项的属性

组织如表1所示。

表1 项目属性矩阵A

商品项	属性			
	A_1	A_2	\dots	A_t
1	C_{11}	C_{12}	\dots	C_{1t}
2	C_{21}	C_{22}	\dots	C_{2t}
\vdots	\vdots	\vdots	\vdots	\vdots
n	C_{n1}	C_{n2}	\dots	C_{nt}

表1中 C_{ij} 表示商品项 i 是否具有属性 A_j ,值为1或0。对于某个用户,可以将其已评价过所有项目投影到相应属性上,以此衡量该用户对不同属性的偏好程度。若项目 i 具有属性 A_k ,投影值为1,否则为0。设 $I(u)$ 代表用户 u 所有已评分项目集合,则用户 u 对项目第 k 个属性的偏好度 $P(u, A_k)$ 可定义如下:

$$P(u, A_k) = \left(\sum_{i \in I(u)} C_{ik} \right) / |I(u)| \quad (9)$$

2.4 改进算法描述

将上述基于用户特征所建立的用户兴趣度、权威度和属性偏好度引入传统单纯依赖用户评分的协同过滤算法中,提出本文的新算法,具体描述如下:

算法1 基于用户特征和项目属性协同过滤算法。

输入:目标用户 T 、项目评分矩阵 R 、项目属性矩阵 A 、邻居数目 k ;

输出:用户 T 的top- N 推荐集。

步骤1 从矩阵 R 中检索出系统的所有用户、项目及目标用户 T 的已评分项目集,分别记为 U_m, I_n, I_T 。

步骤2 采用式(4)中用户兴趣度 $I(u, i)$ 对传统Pearson相关系数法进行加权,改进后的相似性计算方法如式(10)所示。对任一用户 $u \in U_m (u \neq T)$,按此公式计算它与目标用户 T 的相似度,并选择其中相似度较大的前 k 个用户作为目标用户 T 的最近邻居集 $N_T = \{j_1, j_2, \dots, j_k\}$ 。

$$\text{sim}(T, u) = \frac{\sum_{i \in I_{Tu}} (R_{T,i} - \bar{R}_T) \times (R_{u,i} - \bar{R}_u) \times I(u, i)}{\sqrt{\sum_{i \in I_{Tu}} (R_{T,i} - \bar{R}_T)^2} \times \sqrt{\sum_{i \in I_{Tu}} (R_{u,i} - \bar{R}_u)^2}} \quad (10)$$

步骤3 针对目标用户 T 的任一未评分项目 i ,若已有邻居对其评价,则采用式(11),结合邻居用户的评分及信任度预测该项目对目标用户的推荐度 $P_{T,i}$;若项目 i 是新项目,则基于项目属性矩阵 A ,首先计算邻居用户对该项目相关属性的偏好度,然后采用式(12),结合邻居用户的信任度进行预测。

$$P_{T,i} = \bar{U}_T + \frac{\sum_{u \in N_T} T(u) \times (R_{u,i} - \bar{R}_u) \times \text{sim}(T, u)}{\sum_{u \in N_T} |T(u) \times \text{sim}(T, u)|} \quad (11)$$

$$P_{T,i} = \bar{U}_T + \frac{\sum_{u \in N_T} (T(u) \times \sum_{k=1}^t p(u, A_k))}{\sum_{u \in N_T} |T(u)|} \quad (12)$$

步骤4 将预测评分最高的前 N 个项目作为目标用户 T 的top- N 推荐集。

3 实验结果及分析

3.1 数据集和评价标准

实验数据采用美国 GroupLens 项目组提供的 MovieLens

数据集,MovieLens 是一个基于 Web 的研究型推荐系统,用于接收用户对电影的评分并提供相应的推荐列表。数据集中包含由 943 个用户对 1682 部电影的 100 000 个评分,评分范围为 1~5,且每个用户至少对 20 部以上的电影进行了评分。本系统首先将已经标注属性值(0 或 1)的电影导入数据库,选择其中相关的 1 000 条评分作为实验数据集,并将其中的 80% 作为训练集,20% 作为测试集。

为了验证推荐算法的精确度,通常采用平均绝对误差 (Mean Absolute Error, MAE) 进行度量^[11-12],通过计算预测评分与用户实际评分之间的偏差来度量预测的准确性,MAE 的值越小,预测精度越高,推荐质量越好。设经算法预测的 top- N 推荐集的评分为 p_1, p_2, \dots, p_N , 相应的实际评分为 q_1, q_2, \dots, q_N , 则平均绝对偏差 MAE 定义如下:

$$MAE = \frac{\sum_{i=1}^N |P_i - R_i|}{N} \quad (13)$$

3.2 实验设计与结果分析

为了验证本文提出的基于用户特征和项目属性协同过滤算法有效性,设计了两组实验将其与传统单纯基于用户评分的协同过滤算法进行了比较。

首先,测试本文在相似性计算过程中引入的用户兴趣度函数对推荐系统性能的影响,在同一预测方法下,将之与传统相似性计算的余弦相似性 (cosine) 和 Pearson 相关系数方法进行比较,结果如图1所示。

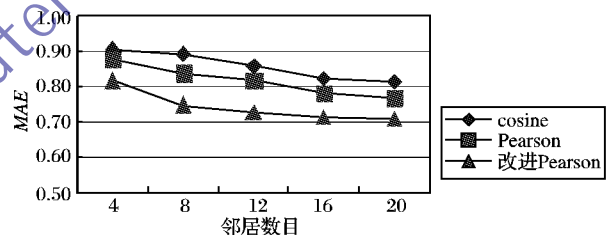


图1 三种相似性度量方法性能比较

从图1中可知,本文所采用的融合了用户兴趣度和评分数据的相似性计算方法,在不同邻居数目下的平均绝对误差 MAE 都明显低于传统方法。可见针对用户兴趣随时间逐步衰减而设计的兴趣度函数,能有效突出用户近期评分数据的重要性,提高兴趣相似邻居用户识别的准确性,所生成的推荐也更加合理有效。

其次,在改进 Person 相似性计算基础上,将融入信任度、属性偏好度等用户特征的本文推荐算法 (User Characteristics and Item Attributes Collaborative Filtering, UCIACF) 与传统基于用户协同过滤推荐算法 (User-Based Collaborative Filtering, UBCF) 的性能进行比较,实验结果如图2所示。

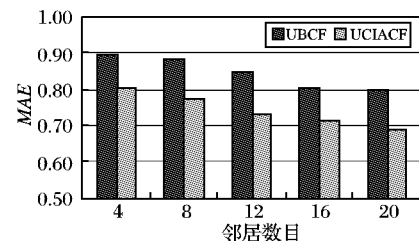


图2 两种推荐算法的精确度比较

可见不同邻居数目下,本文算法的精确度均显著高于传

(下转第 1755 页)

6 结语

信息过滤是人们从海量信息中获取真正所需信息的重要检索技术。但是,信息过滤长期以来受困于精度不高、无法真正实现信息内容语义检索等困难。本文引入本体技术,利用本体对信息进行形式化语义描述,从而使信息与用户要求之间存在可计算的语义基础。同时,我们利用本体实现带约束的语义扩展,将用户未能明确输入的潜在语义扩展出来,并用约束规则对其进行规定,从而避免带来巨大冗余。最终,本文给出了带约束本体语义扩展条件下的信息过滤算法,实现相似度计算下的信息过滤。在以后的工作中,将重点针对形式化约束下的推理展开工作,从而进一步提高语义扩展的精度。

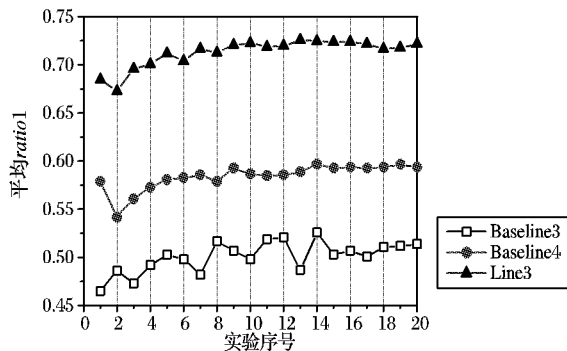


图5 平均过滤效果测试分析(Group3)

参考文献:

- [1] BELKIN J, BRUCE CROFT W. Information filtering and information retrieval: Two sides of the same coin [J]. Communications of the ACM, 1992, 35(12): 29-38.
- [2] BHANDARKAR S M, LUO XING-ZHI. Integrated detection and tracking of multiple faces using particle filtering and optical flow-based elastic matching [J]. Computer Vision and Image Understanding, 2009, 113(6): 708-725.
- [3] GRUBER T R. A translation approach to portable ontology specification [J]. Knowledge Acquisition, 1993, 5(2): 199-220.
- [4] LI KANG, ZHONG ZHEN-YU, LAKSHMISH R. Privacy-aware collaborative spam filtering [J]. IEEE Transactions on Parallel and Distributed Systems, 2009, 20(5): 725-739.
- [5] DESHPANDE A S, TRIANTAPHYLLU E. A greedy randomized adaptive search procedure (GRASP) for inference logical clauses from examples in polynomial time and some extensions [J]. Mathematical and Computer Modelling, 1998, 27(1): 75-99.
- [6] SANCHEZ S N, TRIANTAPHYLLU E, KRAFT D H. A feature mining based approach for the classification of text documents into disjoint classed information [J]. Information Processing and Management, 2002, 38(4): 583-604.
- [7] DREILINGER D, HOWE A E. Experiences with selecting search engine using metasearch [J]. ACM Transactions on Information Systems, 1997, 15(3): 195-222.
- [8] 曾春, 邢春晓, 周立柱. 基于内容过滤的个性化搜索算法 [J]. 软件学报, 2003, 14(5): 999-1004.
- [9] 田范江, 李丛蓉, 王鼎兴. 进化式信息过滤方法研究 [J]. 小型微型计算机系统, 2003, 11(3): 328-333.
- [10] 梁理, 黄樟钦, 侯义斌. 网络信息过滤系统(NFIS)的研究与实现 [J]. 小型微型计算机系统, 2003, 24(2): 195-198.
- [11] 张波, 向阳, 王坚. 一种基于语义可理解的信息过滤算法 [J]. 电子与信息学报, 2010, 32(10): 2324-2330.
- [12] 05]. <http://www.cs.pitt.edu/~mrotaru/comp/rs/Breese%20UAI%201998.pdf>.
- [5] 孙小华. 协同过滤系统的稀疏性与冷启动问题研究 [D]. 杭州: 浙江大学, 2005.
- [6] HERLOCKER J L, KONSTAN J A, BORCHERS A, et al. An algorithmic framework for performing collaborative filtering [C]// Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 1999: 230-237.
- [7] WANG JUN, de VRIES A P, REINDERS M J T. Unifying user-based and item-based collaborative filtering approaches by similarity fusion [C]// Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2006: 501-508.
- [8] 王茜, 王均波. 一种改进的协同过滤推荐算法 [J]. 计算机科学, 2010, 37(6): 226-228.
- [9] 张富国. 用户多兴趣下基于信任的协同过滤算法研究 [J]. 小型微型计算机系统, 2008, 29(8): 1415-1419.
- [10] 李聪, 梁昌勇. 基于属性值偏好矩阵的协同过滤推荐算法 [J]. 情报学报, 2008, 27(6): 884-890.
- [11] HERLOCKER J L, KONSTAN J A, TERVEEN G L, et al. Evaluating collaborative filtering recommender systems [J]. ACM Transactions on Information Systems, 2004, 22(1): 5-53.
- [12] KARYPIS G. Evaluation of item-based top-n recommendation algorithms [C]// Proceedings of the 10th International Conference on Information and Knowledge Management. New York: ACM, 2001: 247-254.

(上接第1750页)

统算法,这是由于我们在预测评分时增加了对邻居用户信任度的度量,体现了不同邻居用户评分对最终目标用户推荐的贡献程度,尤其是引入的属性偏好度函数,使得传统算法中无法解决的新项目得到了有效推荐,从而进一步提高了预测的可靠性和系统的推荐性能。

4 结语

针对传统协同过滤算法中评分数据的极度稀疏及冷启动问题,本文提出了基于用户特征和项目属性的协同过滤推荐算法,通过衡量用户的兴趣变化、受信任程度及对项目不同属性的偏好程度来预测目标用户的评分。相比仅仅依赖用户评分数据寻找最近邻居并进行推荐的传统算法,本文算法能有效改善邻居识别的准确性和预测的可靠性,提高整个系统的推荐质量。

参考文献:

- [1] SCHAFER J B, KONSTAN J A, RIED J. E-commerce recommendation applications [J]. Data Mining and Knowledge Discovery, 2001, 5(1/2): 115-153.
- [2] SARWAR B, KARYPIS G, KONSTAN J, et al. Analysis of recommendation algorithms for e-commerce [C]// ACM Conference on Electronic Commerce. New York: ACM, 2000: 158-167.
- [3] HERLOCKER J L, KONSTAN J A, RIEDL T J. Empirical analysis of design choices in neighborhood-based collaborative filtering algorithms [J]. Information Retrieval, 2002, 5(4): 287-310.
- [4] BREESE J, HECHERMAN D, KADIE C. Empirical analysis of predictive algorithms for collaborative filtering [EB/OL]. [2010-10-