项目子相似度融合的协同过滤推荐算法①

毕孝儒

(四川外国语大学重庆南方翻译学院 管理学院, 重庆 401120)

摘 要: 针对用户评分数据稀疏性和项目最近邻寻找的不准确性问题, 提出了一种项目子相似度融合的协同过滤推荐算法. 该算法根据目标用户每一属性取值, 选取与该属性值一致的用户作为用户子空间, 并在此空间上计算目标项目与其他项目之间的相似度(称其为项目子相似度). 在此基础上, 以项目子相似度为依据选取目标项目的 K 最近邻, 计算其预测评分; 最后对用户不同属性上的预测评分进行加权求和, 得到目标项目的最终评分. 实验结果表明, 该算法能准确地选取目标项目的最近邻, 明显改善了推荐质量.

关键词: 协同过滤; 项目子相似度; 用户属性权值

Collaboration Filtering Recommendation Algorithm of Sub-Similarity Integration between Items

BI Xiao-Ru,

(School of Management, Chongqing Nanfang Translators College of University SISU, Chongqing 401120, China)

Abstract: Aiming at such the problems of sparse data and non-currency to select the nearest neighbors, a collaborative filtering recommendation algorithm of sub-similarity integration between items is proposed in the paper. According to every attribute value of the target user, the users whose attribute value is the same as target user's are selected as user's sub-space, similarity(sub-similarity of items) between the target item and others in the user's sum-space is calculated. Based on it, according to sub-similarity of items, k-nearest-neighbors are selected to calculate it's prediction value. Finally, weighted sum of prediction value of user's attributes is calculated to get final prediction value of the target item. Experimental result shows that the algorithm can select nearest neighbors of target item correctly and improve recommendation quality of spare data.

Key words: collaborative filtering; sub-similarity of items; weighted value of user's attributes

随着互联网上信息的日益增长和用户个性化需求的提高, 推荐系统的应用日益广泛,成为电子商务、社会网络、视频/音乐点播等主流 Web 2.0 服务的核心技术, 而协同过滤(Collaborative Filtering)是推荐系统所采用的最为重要的技术之一. 当前, 协同过滤技术主要分为三类: 基于用户的协同过滤[1]、基于模型的协同过滤和基于项目的协同过滤[2]. 其中, 基于项目的协同过滤根据用户对相似项目的评分, 预测用户对目标条目的评分[3], 因而目标项目邻居的选择是提高推荐质量的关键. 传统的基于项目的协同过滤算法在采用修正余弦相似性和相关相似性等方法计算目标项目与其他项目之间的相似度时, 将每个共评用户(对两个项目共同评分的用户)的评分视为同等重要, 没有考虑共评用户与目标用户属性取值的差异对项目相似度的影响, 致使针对不同目标用户的同一个目标项目的邻居集都相同. 然而, 现实中目标项目的邻居集应随着目标用户的不同而动态变化. 比如, 目标用户的职业是教师, 则受教师喜好的项目应该为目标项目的优先选择邻居.

基于上述分析, 提出了一种项目子相似度融合的协同过滤推荐(Collaborative Filtering Recommendation Algorithm of Sub-similarity Integration between Items,

① 收稿时间:2014-04-25;收到修改稿时间:2014-05-16

SI-CF)算法. 实验结果表明, 该算法可以准确地获取目标项目的最近邻居, 有效地缓解了数据稀疏性问题.

1 传统的基于项目的协同过滤推荐算法

传统的基于项目的协同过滤推荐算法可以分为三个阶段: 第一、根据用户的历史评分记录计算项目之间的评分相似性度; 第二、从与目标项目相似度最高的项目中选取个作为最近邻, 根据这些最近邻项目的实际评分预测用户对目标项目评分; 第三、选取预测评分最高的若干个项目作为推荐结果.

1.1 项目相似度计算

项目相似性度量主要有两种计算方法: 修正余弦相似性和Pearson相关相似性. 设用户对项目的历史评分记录中, 用户集合为 $U=\{U_1,U_2,...,U_m\}$, 项目集合为 $I=\{I_1,I_2,...,I_n\}$, 用户对项目评分数据矩阵表示为 $R=\{r_{i,j}\}_{m\times n}$, 其中, m、n 分别表示用户和项目的个数, 对项目 I_i 评过分的用户集合表示为 U_{I_i}, $\overline{r_{*,i}}$ 为用户集 U_{I_i} 对项目 I_i 评分的平均值, 则项目修正余弦相似度计算方法如式(1)所示:

$$sim(I_i,I_j)=\frac{\sum_{k\in U_{I_i}\cap U_{I_j}}(r_{k,i}-\overline{r_{*,i}})(r_{k,j}-\overline{r_{*,j}})}{\sqrt{\sum_{k\in U_{I_i}}(r_{k,i}-\overline{r_{*,i}})^2}\sqrt{\sum_{k\in U_{I_j}}(r_{k,j}-\overline{r_{*,j}})^2}} \tag{1}$$

项目相关相似度方法如式(2)所示:

$$sim(I_i,I_j)=\frac{\sum_{k\in U_{I_i}\cap U_{I_j}}(r_{k,i}-\overline{r_{*,i}})(r_{k,j}-\overline{r_{*,j}})}{\sqrt{\sum_{k\in U_{I_i}\cap U_{I_j}}(r_{k,i}-\overline{r_{*,i}})^2}\sqrt{\sum_{k\in U_{I_i}\cap U_{I_j}}(r_{k,j}-\overline{r_{*,j}})^2}} \tag{2}$$

1.2 评分预测

根据项目间相似度大小可以获取目标项目的 K 个最近邻居集, 并将其相似性作为权重预测用户对目标项目的评分, 式(3)给出了用户 U_i 对目标项目 I_j 的预测评分计算方法.

$$P_{U_i,I_j}=\overline{r_j}+\frac{\sum_{v\in K_{I_j}}sim(I_j,v)(r_{i,v}-\overline{r_j})}{\sum_{v\in K_{I_j}}(|sim(I_j,v)|)} \tag{3}$$

式(3)中, $\overline{r_j}$ 表示对项目 I_j 所有评分的平均值, K_{I_j} 为项目 I_j 的 K 最近邻居集.

2 项目子相似度融合的协同过滤推荐算法

传统的基于项目的协同过滤推荐算法仅依靠用户的单一评分相似度计算某一用户对目标项目的预测评分, 并未考虑未评分用户的属性特征. 因而, 不能准确反映目标项目与其选取的近邻项目在用户某一属性下的相似程度. 以表 1 描述的用户对不同类别电影项目的评分为例, 若推荐系统需要预测用户 U_3 对电影项目 I_t 的评分, 则在整个用户空间 $\{U_1,U_2,...,U_{12}\}$ 上, 利用式(2)相关相似性方法计算目标项目 I_t 与其他两项目 I_p, I_q 之间相似度分别是为 0.7627、0.8742, 则在预测评分时, 预测公式分配给项目 I_q 的权重将会大于项目 I_p. 但表 1 显示对项目 I_t 评分的用户职业属性是教师, 因此选择职业是教师的用户给出的评分, 以计算项目 I_t 的预测评分更为合理, 即在职业属性为教师用户的子空间 $\{U_1,U_2,...,U_{12}\}$ 上计算目标项目 I_t 与其他两项目 I_p, I_q 之间相似度更符合实际, 其相似度分别是为 0.9702、-0.1584. 则相比较项目 I_q, 项目 I_p 与目标项目 I_t 在教师用户子空间上更为相似. 因此在预测项目 I_t 时, 分配给项目 I_p 权重应大于项目 I_q.

表 1　用户评分数据

		I_p	I_q	I_t
教师	U_1	4	3	4
	U_2			
	U_3	2	2	?
	U_4	3		3
	U_5		3	1
医生	U_6	5	2	
	U_7			
	U_8	3	5	5
程序员	U_9			3
	U_{10}			3
	U_{11}		2	2

为了后续方便分析, 设集合 $A=\{a_1,a_2,...,a_w\}$ 表示用户属性集合, 其中, $a_i(1\le a_i\le w)$ 表示第 i 个属性, 集合 $V=\{v_1,v_2,...,v_z\}$ 表示用户属性集合 A 的所有不同取值的集合, 则对于任意两个项目关于集合 T 的各个不同属性值的 z 个相似度表示为 $sim(I_i,I_j,v_1)$, $sim(I_i,I_j,v_2),...,sim(I_i,I_j,v_z)$.

2.1 项目间子相似度计算

基于前述分析的传统的基于项目的协同过滤中项目相似性度量方法的不足, SI-CF 算法给出一种结合用户属性取值的项目间相似度计算方法. 若 v_j 表示第

$j(1\le j\le z)$ 个属性取值, 若 $U_{p,q}^{v_j}$ 表示对项目 I_p, I_q 评过分的用户的子属性取值为 v_j 的用户集合, $\overline{r_p^{v_j}}$, $\overline{r_q^{v_j}}$ 分别表示子属性取值为 v_j 的用户对项目 I_p, I_q 评分的平均值, 则定义项目 I_p, I_q 关于用户属性值 v_j 的修正余弦相似度计算方法如式(4)所示:

$$sim(I_p,I_q,v_j)=\frac{\sum_{x\in U_{p,q}^{v_j}}(r_{p,x}-\overline{r_p^{v_j}})(r_{q,x}-\overline{r_q^{v_j}})}{\sqrt{\sum_{x\in U_p^{v_j}}(r_{p,x}-\overline{r_p^{v_j}})^2}\sqrt{\sum_{x\in U_p^{v_j}}(r_{q,x}-\overline{r_q^{v_j}})^2}} \tag{4}$$

相应地, 项目 I_p, I_q 关于用户属性值 v_j 的相关相似性度公式如式(5)所示:

$$sim(I_p,I_q,v_j)=\frac{\sum_{x\in U_{p,q}^{v_j}}(r_{p,x}-\overline{r_p^{v_j}})(r_{q,x}-\overline{r_q^{v_j}})}{\sqrt{\sum_{x\in U_{p,q}^{v_j}}(r_{p,x}-\overline{r_p^{v_j}})^2}\sqrt{\sum_{x\in U_{p,q}^{v_j}}(r_{q,x}-\overline{r_q^{v_j}})^2}} \tag{5}$$

由式(5)可知, 通过对两项目关于用户不同属性值的评分记录的计算, 可以获取项目间对于不同用户属性值的各自独立的评分相似度.

2.2 基于项目间子相似度的预测评分计算

考虑到项目之间对于用户的不同属性值有不同的相似度, 在计算用户对目标项目的预测评分时, 首先根据对目标项目预测评分的目标用户的每一属性值, 从其余项目中中选取 K 个与目标项目关于该属性值的相似度最高的项目作为最近邻, 以分别计算目标项目关于目标用户不同属性值的多个预测评分, 最后取其加权平均值作为最终预测评分. 若对目标项目 I_j 预测评分的目标用户 U_i 的 l 个子属性的取值为集合 $V=\{v_{i1},v_{i2},...,v_{il}\}$, 项目 I_j 关于用户 U_i 属性值 $v_{is}(1\le s\le l)$ 的 K 最近邻项目集合为 $N_{I_j}^{v_{is}}$, 则依据式(6), U_i 对 I_j 关于属性值 $v_{is}(1\le s\le l)$ 的预测评分为

$$P(U_i,I_j,v_{is})=\overline{r_j^{v_{is}}}+\frac{\sum_{I'\in N_{I_j}^{v_{is}}}Sim(I_i,I',v_{is})(r_{I',I_j}-\overline{r_j^{v_{is}}})}{\sum_{I'\in N_{I_j}^{v_{is}}}Sim(I_i,I',v_{is})} \tag{6}$$

类似地, 分别计算得到 l 个预测评分 $P(U_i,I_j,v_{i1})$, $P(U_i,I_j,v_{i2}),...,P(U_i,I_j,v_{il})$, 并取其加权平均值作为最后预测评分, 如式(7)所示:

$$P(U_i,I_j)=\sum_{s=1}^{l}\beta_s P(U_i,I_j,v_{is}) \tag{7}$$

上式中, β_k 是权重因子, 且有 $\sum\beta_s=1$, 算法采用 SVDFeature 工具, 通过依次添加用户属性特征的方式, 计算各个属性特征对预测结果的影响程度, 依此计算出各权重的因子所占比例.

2.3 算法描述

算法 1 项目间子相似度计算

输入: 用户集合 U, 项目集合 I, 用户属性取值集合 V, 评分矩阵 $R_{U\times I}$

输出: 项目间子相似度矩阵 $Sim_{I\times I\times V}$

1) *for* $k=1$: $|I|$

2) *for* $j=k+1$: $|I|$

3) *for* $m=1$: $|V|$

4) 根据式(4)或式(5), 计算项目 I_k, I_j 关于属性值为 v_m 的相似度 $sim(I_k,I_j,v_m)$;

5) *end*

6) *end*

7) *end*

算法 2 预测评分计算

输入: 目标用户 U_i 的各子属性的取值集合 V_i, 选取的近邻数目 K, 项目间子相似度矩阵 $Sim_{I\times I\times V}$

输出: 目标用户 U_i 对目标项目 I_j 的评分 $r_{i,j}$

1) *for* $q=1$: $|V_i|$

2) 从项目间子相似度矩阵 $Sim_{I\times I\times V}$ 中选择 K 个与项目 I_j 关于属性值 $v_q\in V_i$ 的相似度最高的项目组成的最近邻集 $N_{I_j}^{v_q}$;

3) 根据式(6), 计算用户 U_i 对项目 I_j 关于属性值 v_x 的预测评分 $P(U_i,I_j,v_q)$;

4) 根据式(7), 计算用户 U_i 对项目 I_j 的最终预测评分 $r_{i,j}$.

5) *end*

3 实验与分析

3.1 实验数据集及评价准则

实验采用 GroupLens 研究小组提供的 MovieLens (http://movieslens.umn.edu)数据集, 它包括 943 个用户对 1 682 个项目的 10 万条投票记录,稀疏等级为 0.9370. 其中, 用户属性有年龄、性别、邮编和职业; 实验采用平均绝对偏差 MAE 作为算法性能评价准则.

3.2 实验环境

实验硬件环境为 Intel inside CORE-i5 系列 CPU、2.2GHz 主频、2GB 内存; 实验软件环境为 Windows7 操作系统、Microsoft VS 2008、SQL Server 2008 数据库.

3.3 用户属性相似度权重因子确定

实验将MovieLens数据集的用户属性(年龄、性别、职业和邮编)依次加入 SVDFeature 的 Group 特征集中, 通过比较 MAE 取值以确定每个属性特征的权重, 结果如图 1 所示. 其中, 用户职业分类判断依据是我国 1999 年 5 月颁布的《中华人民共和国职业分类大典》.

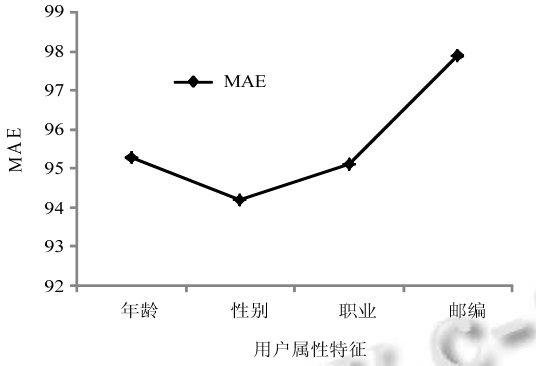

图 1　用户不同属性的 MAE 值

从图 1 可知, 性别对预测结果的影响最大, 年龄次之, 邮编最小. 依据上述实验结果, 分别设置性别、年龄、职业和邮编的权重为 0.45、0.25、0.20、0.1.

3.4 仿真实验结果与分析

实验将数据按 80%和 20%的比例划分为训练集和测试集, 在分别采用修正余弦相似性和相关相似性作为项目相似性度量方法的基础上, 将文中提出的 IS-CF 算法与传统基于项目的协同过滤算法 I-CF、文献[4]基于项目兴趣度的协同过滤算法 II-CF、文献[5]基于项目属性和云填充的协同过滤推荐算法 IACF-CF、文献[6]融合争议度特征的协同过滤推荐算法 CFC-CF 进行了对比实验.

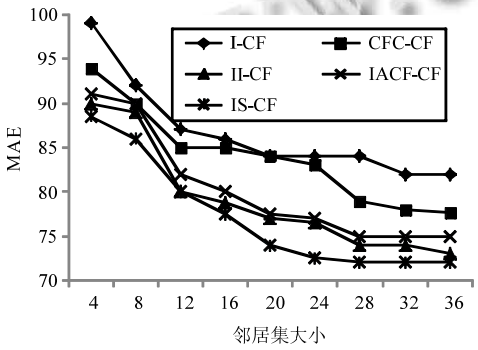

图 2　修正余弦相似性方法方法下的 MAE 值

由图 2、图 3 可知, 无论采用修正余弦相似性还是相关相似性度量方法, 在邻居数目相同的前提下, IS-CF 算法的 MAE 值均小于其他算法. 同时, 随着邻居数目的增加, 上述各种算法的 MAE 下降速度明显减缓, 而邻居数目的增加将增加额外的运算量, 因而在保证推荐质量的前提下, 应尽量选择较少邻居数目.

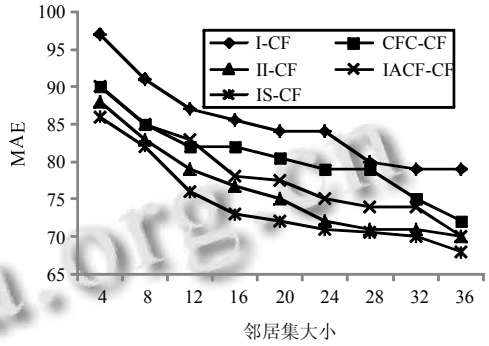

图 3　相关相似性方法下的 MAE 值

4　结语

文中分析了传统的基于项目的协同过滤推荐算法存在的不足, 提出了基于项目间局部相似度融合的协同过滤推荐算法. 实验结果表明, 该算法能准确地获取用户兴趣最近邻, 有效改善了推荐质量. 如何优化算法, 降低项目间局部相似度的计算量, 是下一步研究的主要内容.

参考文献

1 Breese J, Hecherman D, Kadie C. Empirical analysis of predictive algorithm for collaborative filtering. Proc. of the 14th Conference on Uncertainty in Artificial Intelligence. 1998. 43–52.

2 Sarwar B, Karyp IG, Konstan TJ, et al. Item based collaborative filtering recommendation algorithm. Proc. of the 10th International Conference on World Wide Web. Hong Kong. 2001. 285–295.

3 邓爱林,朱扬勇,施伯乐.基于项目评分预测的协同过滤推荐算法.软件学报,2003,14(9):1621–1628.

4 孙光明,王硕.基于项目兴趣度的协同过滤新算法.计算机应用研究,2013,30(12).

5 邓爱林,朱扬勇,施伯乐.基于项目属性和云填充的协同过滤推荐算法.计算机应用,2012,32(3):658–660.

6 张学胜,陈超,张迎峰.融合争议度特征的协同过滤推荐算法.小型微型计算机系统,2012,4(4).