CN 43-1258/TP
ISSN 1007-130X

计算机工程与科学
Computer Engineering & Science

第 35 卷第 5 期 2013 年 5 月
Vol. 35, No. 5, May. 2013

文章编号:1007-130X(2013)05-0161-05

一种基于用户相似性的协同过滤推荐算法*

程　飞,贾彩燕
(北京交通大学计算机与信息技术学院,北京 100044)

摘　要:个性化推荐技术研究用户行为,分析用户兴趣,主动为用户推荐合适的资源,较好地解决了互联网信息日益庞大与用户需求之间的矛盾。协同过滤算法中,基于邻居的方法和基于潜在因子的方法是目前应用于推荐系统最成功的技术。前者虽然简单易行,但精度有待提高;后者精度较高,但模型复杂,参数难以学习。提出了一种改进的基于用户相似性的协同过滤算法,通过修正用户相似性的度量方法,产生更合理的用户邻居,实现对用户的评分推荐。实验结果表明,所提出的算法相比基于潜在因子的方法简单易行;同时,相比基于邻居的方法,在一定程度上提高了推荐的精度。

关键词:推荐系统;协同过滤;用户相似性;用户邻居
中图分类号:TP391　　**文献标志码**:A
doi:10.3969/j.issn.1007-130X.2013.05.026

An improved collaborative filtering algorithm based on user similarity

CHENG Fei, JIA Cai-yan
(School of Computer Science and Information Technology, Beijing Jiaotong University, Beijing 100044, China)

Abstract: The personalized recommendation technique addresses studying the behaviors of individual users, analyzing what they are interested in and recommending suitable resources to them. In other words, the personalized recommendation technique is a better solution to the contradiction between the requirements of users and the explosive information on the Internet. Collaborative filtering algorithms based on neighborhood approach and potential factors are the most successful techniques in the recommendation system. Although the former is easy to implement, its accuracy needs to be improved. Meanwhile, the latter has high precision, but it is complex and the parameters are difficult to learn. Therefore, in the paper, an improved collaborative filtering algorithm based on user similarity is proposed. Through adjusting the measure method of user similarity, it can generate more reasonable user neighbors and recommend the users according to their scores. Experimental results show that the algorithm proposed in this paper is easier to implement than the algorithm based on the potential factors. Besides, compared with the algorithm based on the neighborhood approach, our proposal, to some extent, improves the accuracy.

Key words: recommendation system; collaborative filtering; user similarity; user neighbor

* **收稿日期**:2012-05-15;**修回日期**:2012-09-21
基金项目:国家自然科学基金资助项目(60905029);中央高校基本科研业务费专项资金资助项目;北京市自然科学基金资助项目(4112046)
通讯地址:100044 北京市北京交通大学计算机与信息技术学院
Address: School of Computer Science and Information Technology, Beijing Jiaotong University, Beijing 100044, P. R. China

1 引言

目前,很多大型的电子商务网站为了获得更好的经济效益,通常会给用户推荐合适的资源,以提高用户的使用满意度,如 Amazon、Ebay、Alibaba、Yahoo 等网站都使用了各种形式的推荐系统。在一些商业公司的赞助下,许多计算机研究机构也举办了一些相关的竞赛,如 ACM KDD CUP 2011 Contest、Netflix Prize Competition(2009 年)等。随着用户数量、用户信息量、网络资源数量的增多,推荐系统的设计和实现面临更大的挑战[1]。

为了实现推荐的功能,系统需要根据已有用户对资源的评价,来预测用户对未评价资源的喜爱程度。目前主要有两类方法来解决这样的问题:基于邻居[2,3](Neighborhood Approach)的方法和基于潜在因子模型[4~6](Latent Factor Models)的方法。

基于邻居的方法比较直观,容易理解。这类方法使用统计技术寻找与目标用户有相同或相似兴趣偏好的邻居,根据邻居用户的评分来预测目标用户对资源的评分值,选取预测评分最高的前 N 个资源作为推荐集反馈给目标用户。它的中心思想是有相同兴趣或偏好的用户往往会对同样的资源感兴趣,这也非常符合人们的心理。这类方法的核心是要准确计算目标用户的邻居,也就是用户相似性,所以也称为基于用户(User-Based)的协同过滤方法。类似地,可以考虑资源之间的相似性,使用目标用户评价过的资源集合来预测用户可能感兴趣的其它资源,这类方法称为基于资源(Item-Based)的协同过滤方法。

基于潜在因子模型的方法将用户和资源的特征同时映射到相同的潜在因子空间(Latent Factor Space),以使得它们可以直接比较。这类方法假设用户对资源的评分是多项特征的加权和。例如,当评价的资源是电影时,特征可以是电影的分类,是喜剧或是文艺剧;也可以是电影所适合的观众级别,儿童电影或是大众电影等。但是,在这类模型中,这样的特征并非都是可解释的,也就是说特征不是人为指定的,而是通过机器学习的方法所得到的潜在特征。这类方法中,比较典型的算法有 MFITR[7]、SVD[8]、SVD++[9]等。

邻居模型的方法有效地计算局部的近邻关系,用与目标用户最相近的邻居的行为来估计预测用户的行为,计算复杂度低,直观易理解,能较好地反映用户行为。但是,它没有考虑到全局的用户关系,效果往往不如因子分解模型好。本文提出一种改进的相似性度量方法,更好地体现用户之间的关系,以减小评分的误差,提高推荐的精度。

2 基于用户的协同过滤推荐算法

Sarwar[10,11]等人将基于用户的协同过滤推荐算法分为三个阶段:表示(Representation)、邻居用户形成(Neighborhood Formation)和推荐生成(Recommendation Generation)。

2.1 表示

协同过滤算法通常采用用户-项目评分矩阵 $\boldsymbol{R}(m,n)$ 表示用户评分信息,如表 1 所示。$\boldsymbol{R}(m,n)$ 是一个 $m\times n$ 阶矩阵,其中,m 行表示 m 个用户,n 列表示 n 个资源,R_{ij} 表示用户 i 对资源 j 的评分值。用户对资源的评分可以采用二进制,例如 1 表示喜欢,0 表示不喜欢。同样可以是 5 分制、10 分制或其他度量方式的评分,评分的高低代表用户对资源的喜爱程度的高低。表 1 是用户对资源的评分矩阵。

Table 1 User-item score matrix $\boldsymbol{R}(m,n)$

表 1 用户-资源评分矩阵 $\boldsymbol{R}(m,n)$

	$item_1$	…	$item_j$	…	$item_n$
$user_1$	R_{11}	…	R_{1j}	…	R_{1n}
⋮	⋮		⋮		⋮
$user_i$	R_{i1}	…	R_{ij}	…	R_{in}
⋮	⋮		⋮		⋮
$user_m$	R_{m1}	…	R_{mj}	…	R_{mn}

2.2 邻居用户形成

邻居用户的形成是基于用户的协同过滤推荐算法中最为关键的步骤。对于目标用户 u,我们需要搜索出与它最相近的用户集合 $U=\{u_1,u_2,\cdots,u_k,\cdots,u_k\}$,$u\notin U$ 且 u 与 U 中用户 u_k 之间的相似性 $sim(u,u_k)(1\leqslant k\leqslant K)$ 由大到小排序。

用户相似性度量[12]方法主要有余弦相似性[13](Cosine Similarity)和皮尔逊相关系数[13](Pearson Correlation Coefficient)等。

(1)余弦相似性。

$$sim_{\text{cosine}}(u,v)=\cos(\boldsymbol{u},\boldsymbol{v})=\frac{\boldsymbol{u}\cdot\boldsymbol{v}}{\|\boldsymbol{u}\|_2\times\|\boldsymbol{v}\|_2}=\frac{\sum_{i\in I_{uv}}R_{ui}\cdot R_{vi}}{\sqrt{\sum_{i\in I_{uv}}R_{ui}^2}\sqrt{\sum_{i\in I_{uv}}R_{vi}^2}} \tag{1}$$

其中，$I_{uv}=\{i\in I \mid R_{ui}\neq\emptyset \text{ and } R_{vi}\neq\emptyset\}$（$I$ 表示全部项目空间）表示用户 u、v 的共同评分资源集合，向量 $\boldsymbol{u}$、$\boldsymbol{v}$ 分别表示用户 u、v 在 I_{uv} 上的评分，R_{ui}、R_{vi} 分别表示用户 u、v 对资源 i 的评分。

(2)皮尔逊相关系数。

$$sim_{\text{pearson}}(u,v)=\frac{\sum_{i\in I_{uv}}(R_{ui}-\overline{R_u})\cdot(R_{vi}-\overline{R_v})}{\sqrt{\sum_{i\in I_{uv}}(R_{ui}-\overline{R_u})^2}\sqrt{\sum_{i\in I_{uv}}(R_{vi}-\overline{R_v})^2}} \tag{2}$$

其中，$\overline{R_u}$、$\overline{R_v}$ 分别表示用户 u、v 在 I_{uv} 上的平均得分，即：

$$\overline{R_u}=\frac{1}{|I_{uv}|}\sum_{i\in I_{uv}}R_{ui}$$

$$\overline{R_v}=\frac{1}{|I_{uv}|}\sum_{i\in I_{uv}}R_{vi}$$

2.3 推荐生成

得到目标用户的最近邻集合后，就可以预测目标用户对未评价资源的评分 $\hat{R}_{ui}$[12]。

$$\hat{R}_{ui}=\overline{R_{ul}}+\frac{\sum_{u_k\in U(u)}sim(u,u_k)\cdot(R_{u_ki}-\overline{R_{u_k}})}{\sum_{u_k\in U(u)}sim(u,u_k)} \tag{3}$$

其中，$\overline{R_{ul}}$ 表示用户 u 在全部项目空间 I 上的平均得分，即：

$$\overline{R_{ul}}=\frac{1}{I_u}\sum_{j\in I_u}R_{uj}$$

其中，$I_u=\{j\in I \mid R_{uj}\neq\emptyset\}$。

此时，给目标用户 u 推荐资源时，可以按照它对资源评分的高低排序，推荐前 N 个资源，即 top-N 推荐，这就完成了整个推荐过程。

3 改进的用户相似性度量方法

使用余弦相似性和皮尔逊相关系数的相似性计算两个用户之间相似性时，都是先寻找两个用户共同评分的项集，而忽略了用户未评分的项集，这样就容易导致寻找到的用户邻居不够准确，尤其是在评分数据稀疏的情况下。本文根据 Jaccard 相似性的思想，提出一种改进的用户相似性度量方法。

$$sim_{\text{jaccard}}(u,v)=\frac{|I_u\cap I_v|}{|I_u\cup I_v|} \tag{4}$$

其中，$|I_u\cap I_v|$ 表示用户 u 和用户 v 都评价过的资源的个数，$|I_u\cup I_v|$ 表示用户 u 和用户 v 评价的资源并集的个数。式(4)考虑了用户评价资源对相似性的影响，但未考虑到用户对资源的评分值，也就是说只要用户对资源评分，不管分值的高低，对相似度的计算都无影响。故对式(4)进一步改进，提出加权的 Jaccard 相似性方法。

$$sim_{\text{jaccard2}}(u,v)=\frac{\sum_{i\in I_{uv}}\left(1-\frac{|R_{ui}-R_{vi}|}{r^{\max}}\right)}{\sum_{i\in I_{uv}}\left(1-\frac{|R_{ui}-R_{vi}|}{r^{\max}}\right)+\sum_{j\in I_u\cup I_v-I_{uv}}\left(\frac{r_j^{\max}}{r^{\max}}\right)} \tag{5}$$

其中，$r^{\max}$ 表示用户对资源评分的上限分值，例如，当采用百分制评分时，$r^{\max}$ 为 100。$r_j^{\max}$ 表示资源 j 的最大得分值。式(5)体现了用户评分高低对相似性的影响，即用户对相同资源的评分差值越小，用户之间的相似性就越高。

又考虑到用户之间存在差异性，每个用户有自己的评价标准，有的用户评分普遍较高，而有的用户评分普遍较低。因此，在计算相似性时，对式(5)中的分子做进一步调整：

$$sim_{\text{jaccard3}}(u,v)=\frac{\sum_{i\in I_{uv}}\left(1-\left|\frac{R_{ui}}{r_u^{\max}}-\frac{R_{vi}}{r_v^{\max}}\right|\right)}{\sum_{i\in I_{uv}}\left(1-\left|\frac{R_{ui}}{r_u^{\max}}-\frac{R_{vi}}{r_v^{\max}}\right|\right)+\sum_{j\in I_u\cup I_v-I_{uv}}\left(\frac{r_j^{\max}}{r^{\max}}\right)} \tag{6}$$

其中，$r_u^{\max}$、$r_v^{\max}$ 表示用户 u 和用户 v 的最大评分值。式(6)中，对每个评分值都做了调整，减轻了用户差异对评价带来的影响。

4 实验结果

4.1 数据集

实验数据取自于英国 Glosgow 大学计算机系从 last.fm 网站(www.last.fm)搜集的数据，称为 Last.FM 数据集。该网站是一个提供音乐点播的网站，有非常大的用户群体。该数据集包含3148个用户对 30520 首音乐的点击次数，每个用户平均点击超过 200 首音乐。用户对音乐的点击次数可以看作为用户对音乐的喜爱程度，即点击次数越高，音乐受喜爱程度越高。对 Last.FM 数据集，随机抽取 1/6 的数据作为测试数据，其余的作为训练数据。

4.2 评分数据标准化

由于不同用户对音乐点击次数的差距较大，直接采用点击次数作为对音乐的评分会产生较大误差，故采用标准化的方法将评分范围设定为 0 至

100 分。考虑到用户的差异性,即有些用户点击音乐的次数普遍比较高,而有些则正好相反。举例而言,假如用户 u_1 酷爱音乐,他点击同一首音乐的平均次数在 200 左右,而用户 u_2 收听音乐的时间相对较少,点击同一首音乐的平均次数在 20 次左右。那么,可以认为用户 u_1 收听 200 次音乐和用户 u_2 收听 20 次音乐的喜爱程度相仿。本文按式(7)所示方法标准化评分数据:

$$R_{ui}=\frac{H_{ui}}{H_u^{\max}}\times 100 \tag{7}$$

其中,H_{ui} 表示用户 u 对音乐 i 的点击次数,$H_u^{\max}$ 表示用户 u 点击音乐的最高次数。这样,评分结果标准化为 0～100 分,同时也减轻了用户差异的影响。

4.3 评价指标

本文采用两种指标来评价实验结果:均方根误差 RMSE[9](Root Mean Squared Error)和推荐精度(Precision@N[14])。

$$RMSE=\sqrt{\sum_{(u,i)\in TestSet}(r_{ui}-\hat{r}_{ui})^2/|TestSet|} \tag{8}$$

均方根误差通过计算预测用户评分与实际用户评分之间的偏差来度量预测的精确度,*RMSE* 越小,推荐的质量就越高。

$$Precision@N=\frac{|Prediction(N)\cap Data(test)|}{N} \tag{9}$$

其中,$Precision@N$ 表示对于目标用户,推荐给用户的资源集合,$Data(test)$ 表示测试集中实际应该推荐给用户的资源集合。这种评价方法反映了推荐的精度,值越大,精度越高,推荐的质量就越好。

4.4 实验结果

我们分别实现了 SVD++算法和两种基于用户的协同过滤算法,用户相似性分别使用式(2)和式(6)来度量。

基于用户的协同过滤算法中,邻居数目的选取通常会影响计算结果,邻居数目越多,精度越高,但一般不应超过用户评价资源的平均数。基于用户的协同过滤算法中,各算法的 *RMSE* 比较结果如图 1 所示。

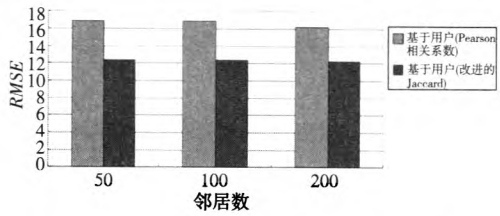

Figure 1 *RMSE* of two algorithms based on user similarity

图 1 两种基于用户相似性算法的 *RMSE* 比较

SVD++算法的 *RMSE* 为 16.15。容易看出,改进算法的预测误差低于其它两种算法。SVD++算法的参数过多,调整较为困难,需要花费大量时间,而且一旦训练数据集发生变化,参数就需要重新调整。SVD++算法使用随机梯度下降的方法来训练模型,梯度大小的选择会对结果造成影响:梯度选择过大,学习的精度不够,预测的结果会不理想,梯度选择过小,又很容易造成"过学习"的情况,而且训练的次数也会明显增多,用时明显增加。基于用户的协同过滤算法则无需考虑参数的调整,实现较为容易。就时间开销而言,本文中改进的用户相似性算法考虑了用户未评分数据,故相比皮尔逊相似性方法计算时间会有所增长。实验结果表明,使用皮尔逊相关系数和本文改进方法计算用户相似性的总耗时分别为 39 分钟和 55 分钟。可见,时间开销的增长并不是很大,仍在可接受的范围之内。本文合理地改进了相似性的度量方法,使传统的基于用户的协同过滤算法也能有较低的预测误差。

各算法的精度比较结果如图 2 所示。从图 2 可以看出,在对前 15 个资源(Precision@15)以及前 20 个资源(Precision@20)的推荐精度上,改进的算法与其它两种算法相比,推荐精度差别不大。但是,在对前 5 个资源(Precision@5)以及前 10 个资源(Precision@10)的推荐精度上,改进算法明显好于另两种算法。在实际应用中,用户往往会关注最优先推荐的资源,而不一定会注意到所有的资源。因此,这种推荐精度的提高,对于实际应用更有价值。

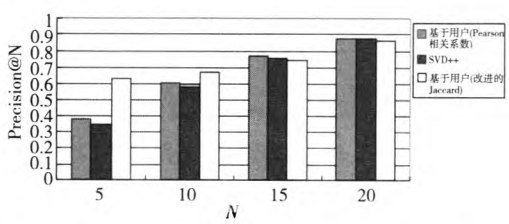

Figure 2 Precision comparison of three algorithms

图 2 各算法的推荐精度对比

5 结束语

基于潜在因子模型的算法体现了全局的评分效果,忽视了局部的特征。传统的基于邻居的算法更多体现了用户的局部特征,更为直观和简单易

行,但精度有待提高。前者参数过多,需要反复调整,一旦有新的用户(或资源)加入,就需要重新训练整个模型,不便于增量式学习。而后者只需计算新增用户(或资源)与其它用户(或资源)之间的相似性就可完成增量学习。本文中的方法在传统的基于用户的协同过滤算法基础上,改进了用户相似性的度量方法,实现了用户对资源评价的预测。改进的用户相似性度量方法考虑了用户差异等行为特点,避免了基于潜在因子模型算法的模型复杂、参数不易学习等缺点;同时,相比传统的基于邻居的方法,改进后的方法在一定程度上提高了推荐质量,有较好的实际应用价值。

参考文献:

[1] Schafer J B, Konstan J A, Riedl J. E-commerce recommendation applications[J]. Data Mining and Knowledge Discovery, 2001, 5(1-2):115-153.

[2] Karypis G. Evaluation of item-based top-N recommendation algorithms[C]// Proc of the 10th International Conference on Information and Knowledge Management, 2001:247-254.

[3] Sarwar B, Karypis G, Konstan J, et al. Item-based collaborative filtering recommendation algorithms[C]// Proc of the 10th International Conference on World Wide Web, 2001: 285-295.

[4] Zhou Yun-hong, Wilkinson D, Schreiber R, et al. Large-scale parallel collaborative filtering for the netflix prize[C]// Proc of the 4th International Conference on Algorithmic Aspects in Information and Management, 2008:337-348.

[5] Agarwal D, Chen Bee-chung, Pang Bo-pang. Personalized recommendation of user comments via factor models[C]// Proc of 2011 Conference on Empirical Methods in Natural Language Processing, 2011:571-582.

[6] Koren Y, Sill J. OrdRec: An ordinal model for predicting personalized item rating distributions[C]// Proc of the 5th ACM Conference on Recommender Systems, 2011:117-124.

[7] Wu Yao, Yan Qiang, Bickson D, et al. Efficient multicore collaborative filtering[C]// Proc of KDD-Cup Workshop, 2011:1.

[8] Takács G, Pilász I, Németh B, et al. Matrix factorization and neighbor based algorithms for the netflix prize problem[C]// Proc of 2008 ACM Conference on Recommender Systems, 2008:267-274.

[9] Koren Y. Factorization meets the neighborhood: A multifaceted collaborative filtering model [C]// Proc of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2008:426-434.

[10] Sarwar B. Sparsity, scalability, and distribution in recommender systems[D]. Minneapolis: University of Minnesota, 2001.

[11] Sarwar B, Karypis G, Konstan O, et al. Analysis of recommendation algorithms for e-commerce[C]// Proc of the 2nd ACM Conference on Electronic Commerce, 2000:158-167.

[12] Adomavicius G, Tuzhilin A. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions[J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(6):734-749.

[13] Ahn Hyung-Jun. A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem[J]. Information Sciences, 2008, 178(1):37-51.

[14] Xu Guan-dong, Gu Yan-hui, Zhang Yan-chun, et al. TOAST: A topic-oriented tag-based recommender system[C]// Proc of the 12th International Conference on Web Information System Engineering, 2011:158-171.

作者简介:

程飞(1987 -),男,浙江建德人,硕士生,研究方向为协同过滤算法和推荐系统。**E-mail**:10120456@bjtu.edu.cn

CHENG Fei, born in 1987, MS candidate, his research interests include collaborative filtering and recommender system.

贾彩燕(1976 -),女,宁夏石嘴人,博士,副教授,研究方向为数据挖掘、机器学习、复杂网络分析和生物信息学。**E-mail**:cyjia@bjtu.edu.cn

JIA Cai-yan, born in 1976; PhD, associate professor, her research interests include data mining, machine learning, complex network analysis, and bioinformatics.

一种基于用户相似性的协同过滤推荐算法

作者： 程飞， 贾彩燕， CHENG Fei， JIA Cai-yan
作者单位： 北京交通大学计算机与信息技术学院,北京,100044
刊名： 计算机工程与科学 ISTIC PKU
英文刊名： Computer Engineering and Science
年，卷(期)： 2013,35(5)

参考文献(14条)

1. Schafer J B. Konstan J A. Riedl J E-commerce recommendation applications 2001(1-2)

2. Karypis G Evaluation of item-based top-N recommendation algorithms 2001

3. Sarwar B. Karypis G. Konstan J Item-based collaborative filtering recommendation algorithms 2001

4. Zhou Yun-hong. Wilkinson D. Schreiber R Large-scale parallel collaborative filtering for the netflix prize 2008

5. Agarwal D. Chen Bee-chung. Pang Bo-pang Personalized recommendation of user comments via factor models 2011

6. Koren Y. Sill J OrdRec:An ordinal model for predicting personalized item rating distributions 2011

7. Wu Yao. Yan Qiang. Bickson D Efficient multicore collaborative filtering 2011

8. Takács G. Pilász I. Németh B Matrix factorization and neighbor based algorithms for the netflix prize problem 2008

9. Koren Y Factorization meets the neighborhood:A multifaceted collaborative filtering model 2008

10. Sarwar B Sparsity,scalability,and distribution in recommender systems 2001

11. Sarwar B. Karypis G. Konstan O Analysis of recommendation algorithms for e-commerce 2000

12. Adomavicius G. Tuzhilin A Toward the next generation of recommender systems:A survey of the state-of-the-art and possible extensions 2005(06)

13. Ahn Hyung-Jun A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem 2008(01)

14. Xu Guan-dong. Gu Yan-hui. Zhang Yan-chun TOAST:A topic-oriented tag-based recommender system 2011