An Improved User-Based Collaborative Filtering Recommendation Algorithm

XIA Jianxun

School of Computer & Information Science, Xiaogan University, Xiaogan, Hubei, China

e-mail: *xia_jx@163.com*

Abstract: User-based collaborative filtering is the most successful and widely used technology in personalized recommender systems to date. The most critical component of the method is the mechanism of finding similarities among users using product ratings data so that products can be recommended based on the similarities. Cosine similarity, correlation similarity and adjusted cosine similarity are three traditional algorithms to measure users' similarity for user-based collaborative filtering recommendation. An improved user-based collaborative filtering recommendation algorithm is proposed. The key point of the proposal is to multiply a weighted value while computing similarity between two users taking advantage of traditional cosine similarity. The weighted value equals to the ratio of the common rating items for two users to the total items. Using the practical data set obtained from GroupLens website and MAE(Mean Absolute Error) metrics for accuracy measure, two traditional algorithms and the proposal algorithm are compared through experiment. Experimental result reveals that the proposed algorithm can yield satisfactory recommendations.

Keywords: personalized recommendation; collaborative filtering; cosine similarity; MAE

一种改进的基于用户的协同过滤推荐算法

夏建勋

孝感学院计算机与信息科学学院，孝感，中国，432000

e-mail: xia_jx@163.com

【摘　要】基于用户的协同过滤技术是到目前为止最为成功和应用最广的推荐技术，其核心是计算两个用户的相似性。计算两个用户的相似性有三种传统算法：余弦相似性、相关相似性和修正的余弦相似性。本文提出了一种改进的基于用户的协同过滤推荐算法，其关键是对两个用户相似性计算方法的改进：在利用传统的余弦相似性求两个用户相似性时增加一个权重值，这个权重值为两个用户共同评分项目数与总项目数的比率。运用 GroupLens 数据集进行改进算法与传统算法的平均绝对误差对比实验，实验结果表明：使用改进的基于用户的协同过滤推荐算法可有效地提高推荐质量。

【关键词】个性化推荐；协同过滤；余弦相似性；平均绝对误差

1 引言

当前，电子商务伴随 Internet 迅猛发展，随之出现了电子商务系统中的信息"超载"现象。海量的商品信息无疑增加了用户购买所需商品的难度，使得用户很难迅速而准确地找到自己真正中意的商品。为此，国内外许多著名电子商务网站诸如亚马逊、eBay、淘宝网等都不同程度地采用推荐系统来进行个性化推荐，以提高用户的点击率，变网站的浏览者为购买者，提高用户购买成功率和交叉销售能力，进而提升网站的美誉度和用户对网站的忠诚度[1]。提供个性化服务已经成为进一步提高网络内容服务质量急需解决的重要课题之一，也是未来网络内容服务的一个发展方向。推荐系统中最为核心和关键的是所采用的推荐技术，它决定了推荐系统性能的好坏。因此，加强对电子商务系统个性化推荐技术研究具有非常重要的实际意义。建立推荐系统有许多技术，例如：协同过滤、贝叶斯网络、聚类分析、关联规则、神经网络等等，其中协同过滤是到应用最早和到目前为止最为成功的推荐技术。协同过滤又分为基于用户的协同过滤和基于项目的协同过滤两类，而基于用户的协同过滤推荐又是最容易理解的一种技术[1]。

2 传统的基于用户的协同过滤推荐算法[2-5]

2.1 算法

基于用户的协同过滤技术是到目前为止实际应用中最为成功的个性化推荐技术，算法的基本思想是将用户感兴趣的项目推荐给具有相同爱好的目标用户。假如目标用户对项目的评分与他的“最近邻居”相似，那么目标用户对某个项目的评分就可以从他的“最近邻居”对某个项目的评分中综合得到。基于用户的协同过滤推荐算法主要包括三个过程：

(1)表示：建立一个用户-项目评分矩阵，描述用户对项目的评价。用户的判断和偏好明确地表示为一个 $m\times n$ 的用户-项目评分矩阵 R，这里 m 是用户数，n 是项目数，$R=\left(R_{i,j}\right)$，元素 $R_{i,j}$ 表示用户 i 对项目 j 的评分。在电子商务推荐系统中，元素 $R_{i,j}$ 既可表示用户是否购买商品(例如 0 或者 1)，也可表示用户对商品的偏好程度(例如评分连续地从 1 到 10 或者离散地为 2、4、6、8、10 等)。用户-项目评分矩阵如表 1 所示。

Table 1. User-item ratings matrix
表 1. 用户-项目评分矩阵

	项目 1	项目 2	…	项目 n
用户 1	$R_{1,1}$	$R_{1,2}$	…	$R_{1,n}$
用户 2	$R_{2,1}$	$R_{2,2}$	…	$R_{2,n}$
…	…	…	…	…
用户 m	$R_{m,1}$	$R_{m,2}$	…	$R_{m,n}$

(2)生成“邻居”：通过计算所有用户对之间的相似性形成“邻居”。计算系统中目标用户与其他所有用户的相似性，以找出 k 个最相似用户集—“最近邻居”。k -“最近邻居”依相似性排序。如图 1 演示了协同过滤中邻居的一种形成过程：当前用户 0 和其它用户之间的相似性被计算，如计算相关相似性。图 1 中与点 0 为中心的 k =5 个最近用户被选择为邻居。

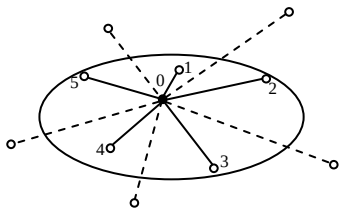

Figure 1. Forming process of neighbors
图 1.邻居的形成过程

(3)产生推荐：通过加权目标用户“邻居”对目标项目的评分产生推荐。根据“最近邻居”集，可计算目标用户对项目的预测评分值，进而产生推荐。假定用户 i 的“最近邻居”集为 S_i，用户 i 对项目 x 的预测评分值为 $P_{i,x}$，用户 i 和用户 j 的相似性为 $sim(i,j)$，用户 i 和用户 j 的平均评分值分别为 $\overline{R_i}$ 和 $\overline{R_j}$，那么有：

$$P_{i,x}=\overline{R_i}+\frac{\Sigma_{j\in S_i} sim(i,j)\times(R_{j,x}-\overline{R_j})}{\Sigma_{j\in S_i}\left(\left|sim(i,j)\right|\right)}$$

2.2 传统的相似性计算方法

计算两个用户之间相似性的传统方法主要有三种：余弦相似性、相关相似性以及修正的余弦相似性，分述如下：

(1)余弦相似性：每个用户的评分作为 n 维项目空间中的一个向量。如果某个用户没有对一个项目评分，则其默认评分值设为 0。两个用户 i 和 j 之间的相似性通过计算两个向量 $\vec{i}$ 和 $\vec{j}$ 夹角的余弦得到，记为 $sim(i,j)$，计算公式如下：

$$sim(i,j)=\frac{\vec{i}\bullet\vec{j}}{\left\|\vec{i}\right\|_2*\left\|\vec{j}\right\|_2}$$

亦即：

$$sim(i,j)=\frac{\sum_{x=1}^{n}(R_{i,x}\times R_{j,x})}{\sqrt{\sum_{x=1}^{n}R_{i,x}^{\ 2}}\sqrt{\sum_{x=1}^{n}R_{j,x}^{\ 2}}}$$

其中：$R_{i,x}$ 为用户 i 对项目 x 的评分值，$R_{j,x}$ 为用户 j 对项目 x 的评分值。

(2)相关相似性：在余弦相似性的计算中没有考虑不同用户之间评分范围的差别。相关相似性通过减去用户对项目平均评分值来克服上述缺陷。两个用户 i 和 j 之间的相似性通过计算 Pearson 相关性得到。假定用户 i 和用户 j 共同评分的项目集合为 I_{ij}（$I_{ij}=I_i\cap I_j$），那么相关相似性计算公式如下：

$$sim(i,j)=\frac{\sum_{x\in I_{ij}}\left(R_{i,x}-\overline{R_i}\right)\left(R_{j,x}-\overline{R_j}\right)}{\sqrt{\sum_{x\in I_{ij}}\left(R_{i,x}-\overline{R_i}\right)^2}\sqrt{\sum_{x\in I_{ij}}\left(R_{j,x}-\overline{R_j}\right)^2}}$$

其中：$R_{i,x}$ 为用户 i 对项目 x 的评分值，$R_{j,x}$ 为用户 j 对项目 x 的评分值，$\overline{R_i}$ 和 $\overline{R_j}$ 分别为用户 i 和用户 j 的平均评分值。

978-1-935068-06-8 © 2009 SciRes.

(3)修正的余弦相似性：在相关相似性计算公式中，如果同时考虑用户 i 和用户 j 的评分项目集合，那么得到修正的余弦相似性计算公式如下：

$$sim(i,j)=\frac{\sum_{x\in I_{ij}}\left(R_{i,x}-\overline{R_i}\right)\left(R_{j,x}-\overline{R_j}\right)}{\sqrt{\sum_{x\in I_i}\left(R_{i,x}-\overline{R_i}\right)^2}\sqrt{\sum_{x\in I_j}\left(R_{j,x}-\overline{R_j}\right)^2}}$$

其中：I_i 和 I_j 分别为用户 i 和用户 j 的评分项目集合。

3 改进的基于用户的协同过滤推荐算法

3.1 算法

在基于用户的协同过滤推荐算法中，最为关键的是如何更准确地寻找相似用户。本文提出的改进算法正是针对传统的余弦相似性计算方法的改进，即在利用传统的余弦相似性计算方法求两个用户相似性时，同时考虑两个用户共同评分项目数与总项目数的比率，把它作为权重加入到传统的余弦相似性计算方法中，这似乎可以增加两个用户的区分度。然后，按照传统的基于用户的协同过滤推荐技术继续进行推荐。本文提出的改进算法中新的余弦相似性计算公式如下：

$$sim(i,j)=\frac{\sum_{x=1}^{n}(R_{i,x}\times R_{j,x})}{\sqrt{\sum_{x=1}^{n}R_{i,x}{}^2}\sqrt{\sum_{x=1}^{n}R_{j,x}{}^2}}\times\frac{\left|I_i\cap I_j\right|}{item\_\max}$$

其中：$R_{i,x}$ 为用户 i 对项目 x 的评价值，$R_{j,x}$ 为用户 j 对项目 x 的评价值，$\left|I_i\cap I_j\right|$ 为两个用户共同评分的项目数，$item\_\max$ 为总项目数。

这样就能保证：只有那些多次参与评分，而且评分项目大体相同的用户，才有可能成为相似用户。而那些只参与了少数几项的评分，或者评分项目差别很大的用户，最终也许不能成为相似用户。上述方法与传统的余弦相似性计算方法相比，由于考虑了两个用户共同评分的项目这个因素，所以，用户之间相似性反映得更客观；与传统的相关相似性计算方法相比，增加了用户之间的区分度，同时并没有减弱用户对不同项目之间评价的差异性；与传统的修正的相似性计算方法相比，用户对不同项目之间评价的差异性仍然保持，但传统的修正的相似性计算方法将一个用户对项目的评分值减去该用户对项目的平均评价值后，用户对不同项目之间评价的差异性就消失了。因此其合理性及准确性明显提高。

利用本文提出的新的余弦相似性计算方法得到目标用户的“最近邻居”，下一步需要产生相应的推荐。用户 i 对项目 x 的预测评分 $P_{i,x}$ 可以通过用户对所有已评分项目的平均评分和用户 i 的“最近邻居”集合中每个用户对目标用户的评分得到。假定用户 i 的“最近邻居”集合为 S_i，用户 i 对项目 x 的预测评分值为 $P_{i,x}$，用户 i 和用户 j 的相似性为 $sim(i,j)$，用户 i 和用户 j 的平均评分值分别为 $\overline{R_i}$ 和 $\overline{R_j}$，那么有：

$$P_{i,x}=\overline{R_i}+\frac{\Sigma_{j\in S_i}sim(i,j)\times(R_{j,x}-\overline{R_j})}{\Sigma_{j\in S_i}\left(\left|sim(i,j)\right|\right)}$$

可以动态扩大“最近邻居”集合的取值范围，适当增加最近邻居的数量，保证推荐结果的可靠性。

具体来说，基于新的余弦相似性计算方法的推荐流程主要包括三个步骤：

(1)建立一个“用户-项目”评分矩阵描述用户对项目的评价；

(2)通过新的余弦相似性计算计算目标用户与其他用户之间的相似性并形成目标用户的“邻居”；

(3)根据目标用户的“邻居”对目标用户未评分项目的评价加权得到目标用户未评分项目的预测评分值，并最终形成推荐。

改进的基于用户的协同过滤推荐算法流程描述如下：

输入：“用户-项目”评分矩阵 R，“邻居”个数 N
目标用户：user1
目标项目：item1
输出：预测值—PredictValue，推荐集 Rec

```
Begin
  For each user in all users
    For each item which user1 rate
      If user rate item
        Number_Of_Common_Rating++
      End if
    Next item
    Similar-
ity(user,user1)=Cosine_Similarity(user,user1)×Number_Of
_Common_Rating/item_max
  Next user
  Select top N user from all users order by similarity
DESC
  For i=1 to N
    Sum_Rating=Sum_Rating + Similar-
```

ity(user,user1)×(Rating$_{user,item1}$ - AVG(user))

Sum_Sim= Sum_Sim+ABS(Similarity(user,user1))

Next i

K=1/ Sum_Sim

PredictValue=AVG(user1) + K×Sum_Rating

End

3.2 实验结果与分析

3.2.1 实验数据集

对于协同过滤推荐算法的比较实验，一般选择著名数据集 EachMovie 进行。EachMovie 是原 DEC 公司系统研究中心的一个专用实验数据集，收集了 1996-1997 年中的 18 个月的数据，共有 72916 个用户对 1628 部电影的 2811983 条评价信息，评价采用离散型的等级分（0，2，4，6，8 ,10）。此数据集原先可从 DEC 公司的网站免费获取。但从 2004 年 10 月起，HP 公司（DEC）关闭了 EachMovie 数据集，所以现在无法直接从网站下载该数据集。为了对比分析传统协同过滤推荐算法与改进的协同过滤推荐算法的推荐效果，本文的实验采用的数据集是目前在衡量推荐算法质量中比较常用的 MovieLens(http://www.grouplens.org/)数据集。该数据集由美国 Minnesota 大学的 GroupLens 研究小组创建并维护。目前，GroupLens 站点免费提供三个数据子集，包括：(1)943 个用户对 1682 部电影的 100000 条评分记录；(2)6040 个用户对 3900 部电影的 1 百万条评分记录；(3)71567 个用户对 10681 部电影的 1 千万条评分记录。其中，每个用户至少对 20 部电影进行了评分。评分的范围是 1~5，5 表示“perfect”，而“1”表示“bad”，用户通过对不同电影的不同评分表达了自己的兴趣。

本文实验选取的是第 1 个数据子集，GroupLens 将

Table 2. Correlative data of experimental data sets

表 2. 实验数据集的相关数据

数据集	数据项数	数据密度	训练集与测试集之比
训练集	80000	5.044%	4：1
测试集	20000	1.261%	

此数据子集随机划分为五个训练集和测试集，数据子集中的 80%作为训练集，剩下的 20%作为测试集。分别为 u1.base 和 u1.test；u2.base 和 u2.test；u3.base 和 u3.test；u4.base 和 u4.test；u5.base 和 u5.test，训练集和测试集的记录数分别为 90570 和 19430 条。GroupLens 实验数据集的相关数据如表 2 所示。

3.2.2 评估标准[2-5]

推荐算法性能的评价是推荐算法实验中的一个重要问题，总体上评价推荐可以有两种方法：一是在线评价，二是离线评价。在线评价是在线调查用户对推荐系统的评价，离线评价是用一个已知的数据集来评价推荐系统的性能。统计精度度量方法中的平均绝对误差 MAE 易于理解，是最常用的一种推荐质量度量方法，它计算算法中预测值和实际值的平均绝对误差（MAE）：

$$MAE = \frac{\sum_{i=1}^{N} |p_i - q_i|}{N}$$

像大多数文献的做法一样，本文以预测值和实际值的平均绝对误差 MAE 作来衡量算法精确度，MAE 越小，推荐质量越高。

3.2.3 实验结果

为了检验算法的有效性和推荐质量，将改进的基于余弦相似性的协同过滤推荐算法同传统的基于相关相似性和修正的余弦相似性的协同过滤推荐算法比较。实验中三种不同算法的平均绝对误差统计数据如表 3 所示。

将表 3 中的 MAE 统计数据用图形显示如图 2 所示。

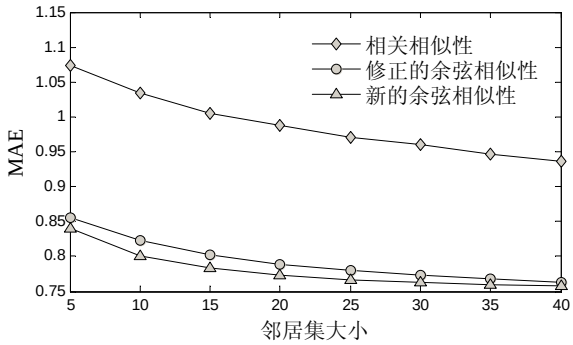

Figure 2. Contrastive curve of MAE

图 2.平均绝对误差对比图

从图 2 可以看出：(1)整体趋势上利用每一种相似性进行协同过滤推荐的结果都是：随着邻居集的增加平均绝对误差减小，即随着邻居集的增加推荐质量提高，但这种平均绝对误差减小的趋势随着邻居集的增加而逐渐变缓。(2)利用本文提出的新的余弦相似性计算方法进行协同过滤推荐，相比利用传统的相关相似性计算方法进行协同过滤推荐的推荐质量有较大改善，而相比利用传统的修正的余弦相似性计算方法进行协同过滤推荐的推荐质量有一定改善。

978-1-935068-06-8 © 2009 SciRes.

Table 3. Statistic of MAE
表 3. 平均绝对误差（MAE）统计数据

相似性	数据集	邻居集大小 5	10	15	20	25	30	35	40
相关相似性	u1	1.0885	1.0459	1.0195	0.9963	0.9819	0.9619	0.9423	0.9317
	u2	1.0883	1.0313	0.9989	0.9794	0.9599	0.9489	0.9417	0.9299
	u3	1.0506	1.0292	1.0033	0.9877	0.9687	0.9654	0.9516	0.9420
	u4	1.0850	1.0496	1.0209	1.0033	0.9870	0.9809	0.9643	0.9501
	u5	1.0531	1.0169	0.9809	0.9727	0.9532	0.9424	0.9363	0.9273
	平均	1.0731	1.0346	1.0047	0.9879	0.9701	0.9599	0.9472	0.9362
修正的余弦相似性	u1	0.8501	0.8234	0.8092	0.7958	0.7843	0.7780	0.7743	0.7692
	u2	0.8627	0.8248	0.8037	0.7906	0.7811	0.7728	0.7674	0.7623
	u3	0.8583	0.8226	0.8039	0.7893	0.7790	0.7712	0.7649	0.7623
	u4	0.8545	0.8238	0.7979	0.7826	0.7749	0.7672	0.7606	0.7570
	u5	0.8564	0.8214	0.7982	0.7889	0.7791	0.7744	0.7689	0.7647
	平均	0.8564	0.8232	0.8026	0.7894	0.7797	0.7727	0.7672	0.7631
新的余弦相似性	u1	0.8282	0.796	0.7839	0.7771	0.7717	0.7682	0.7666	0.7651
	u2	0.8252	0.7886	0.7770	0.7679	0.7620	0.7577	0.7555	0.7546
	u3	0.8599	0.8094	0.7904	0.7783	0.7714	0.7657	0.7613	0.7579
	u4	0.8400	0.8006	0.7807	0.7690	0.7628	0.7592	0.7561	0.7534
	u5	0.8447	0.8044	0.7851	0.774	0.7666	0.7616	0.7583	0.7559
	平均	0.8396	0.7998	0.7834	0.7733	0.7669	0.7625	0.7596	0.7574

4 结论

推荐系统和推荐技术是目前国内外学者的一个研究热点，尤其是协同过滤推荐技术更是得到众多研究者的关注。本文提出了一种改进的基于用户的协同过滤推荐算法：在利用传统的余弦相似性计算两个用户的相似性时，同时考虑两个用户共同评分项目数与总项目数的比率，随后对未评分项目按传统方法进行预测，进而比较了利用相关相似性、修正的余弦相似性和新的余弦相似性三种方法进行预测的预测评分与实际评分的平均绝对误差。结果表明：本文提出的改进的基于余弦相似性的协同过滤推荐算法能有效地提高推荐质量。如何寻找更优化和有效的推荐算法，这是今后进一步的工作。

致谢

在本文完成之际，我首先要感谢在参考文献中提到的各位作者，是他们睿智的思想启迪了我；其次，我还要感谢孝感学院，是她给我提供了从事科学研究的良好环境。

References (参考文献)

[1] Yu Li, Liu Lu, Research on personalized recommendations in E-business[J], *Computer Integrated Manufacturing Systems*, 2004, 10(10), P1306-1313 (Ch).
余力，刘鲁，电子商务个性化推荐研究[J]，计算机集成制造系统，2004, 10(10)，P1306-1313.

[2] Zhao Liang, Hu Naijing, Zhang Shouzhi, Algorithm design for personalization recommendation systems[J], *Journal of Computer Research and Development*, 2002, 39(8), P986-991 (Ch).
赵亮，胡乃静，张守志，个性化推荐算法设计[J]，计算机研究与发展，2002, 39(8)，P986-991.

[3] Zhao Zhi, Shi Bing, An adaptive algorithm for personal recommendation[J], *Journal of Changchun University*, 2005, 15(6), P26-29 (Ch).
赵智，时兵，改进的个性化推荐算法[J]，长春大学学报，2005, 15(6)，P26-29.

[4] Zhou Junfeng, Tang Xian,Guo Jingfeng, An optimized collaborative filtering recommendation algorithm[J], *Journal of Computer Research and Development*, 2004, 41(10), P1842-1847 (Ch).
周军锋，汤显，郭景峰，一种优化的协同过滤推荐算法[J]，计算机研究与发展，2004, 41(10)，P1842-1847.

[5] Xu Jianchao, Wang Hongmei, Improvement on correlation similar collaborative filtering algorithm[J], *Journal of Jilin University(Information science edition)*, 2008, 26(1), P99-105 (Ch).
许建潮，王红梅，改进的协同过滤算法[J]，吉林大学学报(信息科学版)，2008, 26(1)，P99-105.