基于专家动态生成的协同过滤推荐算法

范中磊，贾彭慧

（长安大学 信息工程学院，西安 710064）

**摘 要：**协同过滤算法通过已知评分估计未知评分，是推荐系统的重要组成部分。本文在传统协同过滤算法的基础上，提出一种基于专家动态生成的协同过滤推荐算法, 该算法从普通用户中挖掘专家，为每个用户建立一个专家数据库，并计算目标用户-专家相似度，利用近邻算法预测目标用户的未知评分。在FilmTrust数据集上的实验结果表明，相对于基于用户的协同过滤算法（UBCF）和采用均值中心化改进UBCF的算法（RUBCF）来说，MAE分别降低了45.31%和5.83%， RMSE分别降低了64.89%和10.28%，有效的提高了推荐系统的准确率。

**关键词：**协同过滤；推荐系统；专家动态生成

Expert Dynamic Generation-Based Collaborative Filtering Recommendation Algorithm

FAN Zhong-lei, JIA Peng-hui

(School of Information Engineering, Chang’an University, Xi’an 710064)

**【Abstract】**Collaborative filtering algorithms estimate unknown scores by known scores and are an important part of the recommendation system. Based on the traditional collaborative filtering algorithm, this paper proposes a collaborative filtering recommendation algorithm based on expert dynamic generation. This algorithm mines experts from ordinary users, builds an expert database for each user, and calculates target users-expert similarity. The neighbor algorithm is used to predict the unknown score of the target user. The experimental results on the FilmTrust dataset show that compared to the user-based collaborative filtering algorithm (UBCF) and the algorithm for improving the UBCF with mean centering (RUBCF), the MAE is reduced by 45.31% and 5.38% respectively ,and the RMSE is reduced by 64.89% and 10.28% respectively,effectively improve the accuracy of the recommendation system.

**【Keywords】**Collaborative Filtering; Recommendation System; Expert Dynamic Generation

**中图分类号：**TP399 **文献标识码：**A

**1 前言**

随着信息技术和互联网的迅速发展，人类逐渐进入信息过载的时代[1]，推荐系统应运而生。推荐系统根据用户的需求信息和行为特征为用户提供个性化服务，这在很大程度上解决了信息过载的问题。

根据推荐系统的组成部分，推荐系统可分为基于内容的推荐系统、协同过滤（Collaborative Filtering，简写为CF）推荐系统和混合推荐系统[2]。协同过滤在推荐系统中应用最为广泛，在商业中取得了巨大的成就，其中包括著名的电子商务网站Amazon.com。协同过滤又分为基于用户的协同过滤—UBCF（User-based Collaborative Filtering）和基于商品的协同过滤。本文研究基于用户的协同过滤[3]，它的基本思想是根据相似用户的兴趣爱好和历史行为，为目标用户推荐商品。因此，在协同过滤算法中识别相似用户的能力是提高推荐质量最重要的部分[4]。

在实际应用中，用户历史行为过少会引起数据稀疏性问题，使计算相似度变得十分困难，造成推荐系统的准确率较低。为了解决这一问题，国内外学者进行了大量的研究工作。文献[5]提出均值中心化UBCF算法—RUBCF（User-based Resnick Collaborative Filtering），结果表明该算法能够提供更准确的预测评分。文献[6]将信任因素引入协同过滤推荐算法，提高了推荐系统的准确率。文献[7]首次利用专家数据集计算相似度，改善了推荐质量。文献[8]提出了基于专家用户的可信推荐系统，在普通用户中挖掘出固定的专家数据集，该方法改进了推荐的正确率。

虽然上述文献提到的方法在一定程度上提高了推荐系统的准确率，但仍然存在一些不足之处：1）某些领域的专家用户匮乏，在数据稀缺的情况下无法直接形成专家数据集；2）在普通用户中定义的专家是固定的，并没有针对每个普通用户建立与之相对应的专家数据集。

针对以上问题，本文提出了一种基于专家动态生成的协同过滤推荐算法—EDCF（Expert Dynamic Generation-Based Collaborative Filtering），为每个用户建立一个专家数据库。在FilmTrust数据集上的测试结果表明该算法提高了推荐系统的准确率，证明了它的有效性。

本文其余部分安排如下：第二节介绍EDCF算法设计，包括EDCF的基本思想和算法概述。第三节对专家评估标准进行具体介绍。第四节通过实验对比和分析对EDCF算法进行评估。最后总结全文并对未来工作进行展望。

**2 EDCF算法设计**

在实际生活中，每个领域都有专业知识较为全面的人，称这些人为专家。对普通用户来说，专家的意见往往更具有参考价值。为了提高推荐的准确率，EDCF算法利用专家评估标准，从用户之间的相对交互性、可靠性、趋同性三个方面考虑，为各个用户建立属于自己的专家数据库。

图1给出了EDCF算法的概述图，其中包括4部分：

（1）建立专家数据库：根据专家评估标准为每个用户建立专家数据库；

（2）为目标用户挑选k位近邻专家：在推荐系统领域中，基于近邻的方法[9]是最早使用的方法之一，不同的专家会产生不同的推荐效果，根据为每个用户建立的专家数据库中的专家因子挑选k位近邻专家作为该用户的专家数据集；

（3）计算目标用户与专家的相似度：本文采用了一种改进的余弦相似度计算方法[7]，计算用户u和专家e之间的相似度时，考虑了两个用户共同评定的商品数量，即调整因子，公式如下：

 （1）

其中代表用户u和专家e共同评价商品的集合，表示用户u和专家e共同评定商品的数量，表示用户u评定商品的数量，表示专家e评定商品的数量；

（4）预测目标用户评分：采用Resnick公式[5]的变形，则用户u对商品i的预测评分如式（2）所示：

 （2）

其中，表示用户u对所有商品评分的均值，表示专家e对商品i的评分值，表示专家e对商品评分的均值。



图1 EDCF算法的概述图

**3专家评估标准**

文献[8]中定义了评估专家的标准，利用该标准进行实验后，发现实验结果尚有改进的空间，本文从用户之间的相对交互性、可靠性以及趋同性提出了新的评估专家的标准。

**3.1 相对交互性**

两个用户共同评价商品的数量越多，表示用户之间的交互性越强，计算用户相对交互性的公式如下：

 （3）

其中，表示用户之间的相对交互性，表示用户u和v共同评价商品的数量，表示用户之间共同评价商品数量的最大值。

**3.2相对可靠性**

信任因素在推荐中起到了关键作用，专家理应拥有更多的信任人数，计算用户相对可靠性的公式如下：

 （4）

式中，表示用户之间的相对可靠性，表示用户v拥有信任者的数量，表示用户所拥有信任者数量的最大值。

**3.3相对趋同性**

两个用户对相同商品的评分越接近，他们的差异性越小，计算评分差异的平均值，并对其归一化表示为用户之间的相对趋同性，公式如下：

 （5）

式中是用户之间的相对趋同性，表示用户u对商品i的评分。

**3.4 定义专家因子**

综合用户的相对交互性、相对可靠性、相对趋同性，定义专家因子的公式如下：

 （6）

其中代表各部分所占权重，。

由公式（6）可以得知：对用户而言，专家因子的值越大，说明专家越专业。

**4 实验**

**4.1 实验数据集**

实验采用FilmTrust数据集[10]对算法的可行性进行验证。FilmTrust数据集来自一个基于社交网络的电影推荐系统。该数据集由两部分组成，用户评分集和信任集。

用户评分集中包括1508位用户对2071个商品的35494条评分数据，评分值的范围是[0.5,4.0]，评分值越大，表示用户越喜欢该商品。用户评分数据集的稀疏度为98.8635%。

信任集中包括用户之间的1853条信任关系评分数据，评分值为1或者空，1表示用户的信任评分值，空值表示用户之间没有信任评分。

**4.2 评估标准**

精确率是评估推荐系统过程中最基本的指标之一。本文分别采用平均绝对误差MAE和均方根误差RMSE作为精确率度量标准。

平均绝对误差MAE通过比较预测值与用户实际评分值之间的偏差来衡量预测的准确性。MAE越小，表明推荐结果越精确。假设预测的用户评分集合表示为，对应的实际用户评分集合是，n表示所有评分的数目，则计算平均绝对误差MAE的公式如下：

 （7）

均方根误差RMSE是预测值与真实值偏差的平方与观测次数n比值的平方根，RMSE值越小，表明推荐质量越高。均方根误差RMSE的计算公式如下：

 （8）

**4.3 实验设计与结果分析**

本文分别对EDCF算法、UBCF算法和RUBCF算法进行实验，并从实验结果的平均绝对误差MAE和均方根误差RMSE两方面进行比较和分析。为了提高仿真结果的真实性，实验将Filmtrust数据集中的评分集按照8:2的比例随机分为训练集和测试集。目标用户中有一些用户属于新用户，即从未有过历史记录的用户。对于新用户，实验处理的方法是：首先在测试集中确定新用户所预测的商品，然后在训练集中找出对该商品已进行过评分的用户，最后将这些用户对该商品评分的均值作为新用户的评分。在专家因子的公式中，我们通过枚举法改变权重值来观察MAE和RMSE值的变化，最终发现当时，精确率最高。

（1）合适的近邻专家数量k

为了确定合适的近邻专家数量k，实验采用改进的余弦相似度作为相似性度量方法，分别选取10，20，30，40，50，60，70，80，90，100，150和200作为近邻值，可得到如图2所示的预测结果。



图2 近邻专家数量k

图2的实验结果反映了近邻专家数量k与平均绝对误差MAE和均方根误差RMSE之间的关系。从实验结果的变化趋势来看，MAE和RMSE均随着k的增加呈现先下降后上升的趋势，当近邻专家数量k=100时，平均绝对误差MAE和均方根误差RMSE最低，分别为：0.81976和0.79152。

（2）各种协同过滤算法的比较

图3显示了三种协同过滤算法的比较。



图3 三种协同过滤算法的比较

图3为EDCF算法、UBCF算法和RUBCF算法的平均绝对误差MAE和均方根误差RMSE的对比图。从实验结果可以看出，EDCF算法相对于UBCF算法和RUBCF算法，MAE分别降低了45.31%和5.83%， RMSE分别降低了64.89%和10.28%。

**5结束语**

本文针对目前基于用户的协同过滤算法中存在的推荐精确度不高的问题，提出一种基于专家动态生成的协同过滤算法（EDCF）。在FilmTrust数据集上的实验结果验证了该算法的有效性。EDCF算法对实际生活中的推荐系统有重要的研究意义和价值。尽管本文提出的算法已经显示出它的确可以提高推荐系统的精确度，但尚存在改进的余地。首先，本文在处理新用户时采用了较为简单的方法，在今后的工作中可进一步考虑将新用户与EDCF算法相结合进行研究；其次，该算法在计算相似度的方法上还需继续改进，以得到更好的推荐效果。

**参考文献**

1. Rutkowski A F, Saunders C S. Growing Pains with Information Overload[J]. Computer, 2010, 43(6):96-95.
2. Balabanović M, Shoham Y. Fab: content-based, collaborative recommendation[J]. Communications of the ACM, 1997, 40(3):66-72.
3. Breese J S, Heckerman D, Kadie C. Empirical Analysis of Predictive Algorithms for Collaborative Filtering[J]. Uncertainty in Artificial Intelligence, 2013, 98(7):43-52.
4. Jesús Bobadilla, Ortega F, Hernando A , et al. Improving collaborative filtering recommender system results and performance using genetic algorithms[J]. Knowledge-Based Systems, 2011, 24(8):1310-1316.
5. Resnick P . GroupLens : An open architecture for collaborative filtering of netnews[C]// Proc. CSCW94 : Conference on Computer Supported Cooperative Work, Chapel Hill. Addison-Wesley, 1994.
6. O'Donovan J, Smyth B. Trust in recommender systems[C]//Proceedings of the 10th international conference on Intelligent user interfaces. ACM, 2005: 167-174.
7. Amatriain X , Lathia N , Pujol J M , et al. The wisdom of the few: A collaborative filtering approach based on expert opinions from the web[C]// International Acm Sigir Conference on Research & Development in Information Retrieval. ACM, 2009.
8. 王宁,徐锋.基于专家用户的可信推荐系统[C].//2011全国软件与应用学术会议(NASAC2011)论文集.南京大学,2011:205-210.
9. Desrosiers C, Karypis G. A Comprehensive Survey of Neighborhood-based Recommendation Methods[M]// Recommender Systems Handbook. Springer US, 2015.
10. <https://www.librec.net/datasets.html>

**作者简介：**

范中磊(1970－)，男，博士，教授，研究方向为分布式计算与存储

贾彭慧(1993－)，女，硕士研究生，研究方向为推荐系统

联系人：范中磊

通讯地址：西安市南二环中段 长安大学信息学院

邮政编码：710064

电话：15991853582 或 029-62630043

E-mail：[zlfan@chd.edu.cn](mailto:zlfan@chd.edu.cn)