# 基于专家动态生成的协同过滤推荐算法

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。本文在传统的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，并构建目标用户-专家相似度矩阵，利用近邻算法预测未知评分。在FilmTrust数据集上的实验结果表明，该方法提高了推荐系统的准确率。

**关键词**：协同过滤，推荐系统，专家

## 1 引言

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（CF）推荐系统和混合推荐系统[3]。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[16]。协同过滤又分为基于用户的协同过滤和基于商品的协同过滤。本文只研究基于用户的协同过滤，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[4,5]。

在实际应用中，由于用户历史行为过少而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不高。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[8]提出了一种新的计算相似度的方法，根据不同的目标用户选择不同的相似用户。文献[9]将信任因素引入协同过滤推荐系统模型中，提高了推荐系统的准确率。文献[10]首次利用专家数据集计算相似度，提高了推荐质量。文献[11]提出了基于专家用户的可信推荐系统，该方法提高了推荐的正确率和覆盖率。

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。

考虑到以上问题，本文提出了一种基于专家动态生成的协同过滤推荐算法（简称EDBCF），不仅考虑到从普通用户中挖掘出专家，而且为每个用户建立一个专家数据库，利用近邻性完成专家动态生成。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。

本文第2节提出了专家动态生成模型；第3节是实验；最后是总结和展望。

## 2专家动态生成模型

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。在可直接得到专家数据集的情况下，利用基于专家的最近邻算法通过直接计算用户与专家的相似度完成推荐。然而，一般很难直接得到专家数据集，需要从普通用户中挖掘专家。对于每一个用户而言，专家数据集也在不断变化。为了提高推荐的准确率，本文提出了基于专家动态生成的协同过滤算法（简称EDBCF），如图1所示。该算法包括5部分：（1）建立专家数据库；（2）挑选k位近邻专家；（3）计算目标用户与专家的相似度；（4）预测目标用户评分；（5）为新用户预测评分。



图1：专家动态生成模型图

### 2.1建立专家数据库

每个用户都会对应不同的专家数据集，根据用户之间的相对活跃性、相对信任性、相对差异性为每个用户建立专家数据库。

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，计算用户相对活跃性的公式如下：

 （1）

其中，表示用户的相对活跃性，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。

信任因素在推荐中起到了关键作用[6,7]，专家用户理应拥有更多的信任人数，计算用户可靠性的公式如下：

 （2）

其中，表示用户之间的相对信任性，表示对于用户u来说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。

对用户来说，所对应的专家评分应该与自身评分的差异越小越好，选择两个用户评分偏差的平均值来表示用户之间的相对差异性，公式如下：

 （3）

式中是用户之间的相对差异性，表示用户u对商品i的评分。

综合用户的相对活跃值、相对信任值、相对差异值，定义专家因子的公式如下：

 （4）

其中代表各部分所占权重，。

### 2.2挑选k位近邻专家

在推荐系统领域中，基于近邻的方法[23]是最早使用的方法之一，不同的专家数量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。

### 2.3计算用户-专家的相似度

在协同过滤算法中，计算用户之间的相似度主要有两种方法：余弦相似度和相关相似性。本文只采用余弦相似度进行计算。在基于余弦的方法[12]中，两个用户x和y被视为m维空间中的两个向量，可以通过计算它们之间角度的余弦值来测量两个矢量之间的相似性：

 （5）

其中代表用户x和y共同评价商品的集合。

本文采用了一种改进的计算相似度的算法[10]，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：

 （6）

其中，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。

### 2.4 预测目标用户评分

当目标用户有相应的专家数据库时，采用Resnick公式[17]的变形，则用户u对商品i的预测评分如式（9）所示：

 （7）

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。

### 2.5 新用户评分

目标用户中有一些用户属于新用户，即从未有过打分记录的用户。对于新用户，本文首先找到了新用户所预测的商品，然后找到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。

## 4. 实验

### 4.1．实验数据集

实验采用FilmTrust数据集[21]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，评分值为1~4.5中的任意值，评分值越大，表示用户越喜欢该商品。用户-商品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[18]。

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：

 （8）

均方根误差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，均方根误差也是用来衡量预测值同真实值之间的偏差，偏差越小，表明推荐质量越高。均方根误差RMSE的定义如下：

 （9）

### 4.3 实验设计与结果分析

本实验的目的是将基于专家动态生成的协同过滤推荐算法（EDBCF）分别与传统基于用户的协同过滤算法（UBCF）[15]和采用Resnick公式并基于用户的协同过滤算法（RUBCF）进行比较分析。

实验将Filmtrust数据集中的评分集按照8:2的比例分为训练集和测试集。

1. 合适的专家数量k
2. 各种算法的比较

## 结束语

本文针对目前传统协同过滤算法存在的推荐精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。该算法在普通用户中挖掘专家用户，并为每个用户建立一个专家数据库，在FilmTrust数据集上的实验结果验证了该算法的有效性。该算法相比较其它两种算法来说，精确率最高，而且也大大地减少了计算成本。

尽管本文提出的算法已经显示出它的有效性，但也为未来的改进提供了机会。首先，如何更加有效地从普通用户中挖掘专家；其次，该算法在不同数据集上最优参数选取也需进一步调整；再者，专家用户也可以和各种CF结合使用，以提高推荐的准确性。

**参考文献**

[1 ] C. S. Saunders and A. Rutkowski, "Growing Pains with Information Overload," in Computer, vol. 43, no. , pp. 96, 94-95, 2010. doi:10.1109/MC.2010.171

[2] A. Tuzhilin and G. Adomavicius, "Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions," in IEEE Transactions on Knowledge & Data Engineering, vol. 17, no. , pp. 734-749, 2005.

[3] [Marko Balabanović , Yoav Shoham, Fab: content-based, collaborative recommendation, Communications of the ACM, v.40 n.3, p.66-72, March 1997](https://dl.acm.org/citation.cfm?id=245124) [doi>[10.1145/245108.245124](http://doi.acm.org/10.1145/245108.245124)]

[4] J. Bobadilla, F. Ortega, A. Hernando, J. Alcala, Improving collaborative filtering recommender system results and performance using genetic algorithms,Knowledge-Based Systems 24 (2011) 1310–1316.  
[5] J. Bobadilla, F. Serradilla, J. Bernal, A new collaborative filtering metric that improves the behavior of recommender systems, Knowledge-Based Systems 23 (2010) 520–528.

[6]R. Levien. Attack-resistant trust metrics. Computing with Social Trust,Spring,pp.121-132,2009

[7]C.Ziegler ,S. McNee，J.Konstan,and G.Lausen.Improving recommendation lists through topic diversification. World Wide Web Conference,pp. 22-32,2005

[8] H.J. AhnA new similarity measure for collaborative filtering to alleviate the new user cold-starting problem Information Sciences, 178 (2008), pp. 37-51

[9] [John O'Donovan , Barry Smyth, Trust in recommender systems, Proceedings of the 10th international conference on Intelligent user interfaces, January 10-13, 2005, San Diego, California, USA](https://dl.acm.org/citation.cfm?id=1040870) [doi>[10.1145/1040830.1040870](http://doi.acm.org/10.1145/1040830.1040870)]

[10] [Xavier Amatriain , Neal Lathia , Josep M. Pujol , Haewoon Kwak , Nuria Oliver, The wisdom of the few: a collaborative filtering approach based on expert opinions from the web, Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, July 19-23, 2009, Boston, MA, USA](https://dl.acm.org/citation.cfm?id=1572033) [doi>[10.1145/1571941.1572033](http://doi.acm.org/10.1145/1571941.1572033)]

[11] 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.

[12] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, “Item-Based Collaborative Filtering Recommendation Algorithms,” Proc. 10th Int’l WWW Conf., 2001.

[13] U. Shardanand and P. Maes, “Social Information Filtering:Algorithms for Automating ‘Word of Mouth’,” Proc. Conf. Human Factors in Computing Systems, 1995.

[14] P. Resnick, N. Iakovou, M. Sushak, P. Bergstrom, and J. Riedl,“GroupLens: An Open Architecture for Collaborative Filtering of Netnews,” Proc. 1994 Computer Supported Cooperative Work Conf.,1994.

[15]J. Breese，D.Heckerman，and C. Kadie. Empirical analysis of predictive algorithms for collaborative filtering. Conference on Uncertainty in Atificial Intelligence,1998

[16] G. Linden, B. Smith, and J. York, “Amazon.com Recommendations: Item-to-Item Collaborative Filtering,” IEEE Internet Computing, Jan./Feb. 2003.

[17]Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. Grouplens: An open architecture for collaborative filtering of netnews. In Proceedings of ACM CSCW’94 Conference on Computer Supported Cooperative Work, Sharing Information and Creating Meaning, pages 175–186, 1994.

[18] T. Chai and R. R. Draxler. Root mean square error (RMSE) or mean absolute error (MAE)? –Arguments against avoiding RMSE in the literature. Geoscientific Model Development,7(3)，pp. 1247-1250，2004

[19] Herlocker J., Konstan J., Terveen L. and Riedl J. 2004. Evaluating collaborative filtering recommender systems. ACM Transactions on Information Systems, 22(1), pp. 5–53.

[20] [Mouzhi Ge , Carla Delgado-Battenfeld , Dietmar Jannach, Beyond accuracy: evaluating recommender systems by coverage and serendipity, Proceedings of the fourth ACM conference on Recommender systems, September 26-30, 2010, Barcelona, Spain](https://dl.acm.org/citation.cfm?id=1864761) [doi>[10.1145/1864708.1864761](http://doi.acm.org/10.1145/1864708.1864761)]

[21] <https://www.librec.net/datasets.html>

[22]Golbeck, J. and Hendler, J. 2006. FilmTrust: Movie recommendations using trust in Web-based social networks. In Proceedings of the IEEE Consumer Communications and Networking Conference. Las Vegas, NV.]]

[23] Christian Desrosiers G K . A Comprehensive Survey of Neighborhood-based Recommendation Methods[M]// Recommender Systems Handbook. Springer US, 2015.