综合用户特征及专家信任的协作过滤推荐算法

高发展 黄梦醒 张婷婷
(海南大学信息科学技术学院 海口 570228)

摘 要 协作过滤推荐算法是推荐系统中应用最广泛的算法之一。通过分析传统协作过滤算法中由数据稀疏性导致的推荐精度不高的问题,在基于专家信任的协作过滤推荐算法的基础上,提出了一种综合用户特征及专家信任的协作过滤推荐算法。该算法分析了用户的不同特征,比较了用户与专家的相似度,通过计算用户-专家相似度矩阵,有效降低了数据集的稀疏性,提高了预测的准确性。在MovieLens数据集上的实验结果表明,改进的算法能够有效缓解冷启动问题,明显提高了系统的推荐精度。

关键词 专家信任,用户特征,协作过滤
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.02.014

Collaborative Filtering Recommendation Algorithm Based on User Characteristics and Expert Opinions

GAO Fa-zhan HUANG Meng-xing ZHANG Ting-ting
(College of Information Science & Technology, Hainan University, Haikou 570228, China)

Abstract Collaborative filtering recommendation algorithm is one of the most widely used algorithms in recommender system. After analyzing the low precision problem caused by sparse data in conventional collaboration filtering algorithms, this paper proposed an collaboration filtering algorithm which integrates user characteristics and expert opinions. The algorithm analyzes user characteristics, compares the similarity between users and experts, and then calculates the similarity matrix. Our algorithm reduces the sparsity of dataset and improves the accuracy of prediction. Our experimental results based on the MovieLens dataset show that, by using our algorithm, performance on the cold start problem and relevant accuracy of recommendation has greatly improved.

Keywords Expert opinions, User characteristics, Collaborative filtering

1 引言

随着网络和信息技术的飞速发展,互联网上的信息和知识都处于大爆炸状态,使得用户无法从大量信息中获得对自己有用的信息,这种信息超载是当前互联网时代所面临的一个重大难题。推荐系统是解决信息超载的有效方法之一,其中学术界和工业界应用最广泛的是协作过滤算法[1-2]。它根据用户的兴趣特点和购买行为向用户推荐其感兴趣的信息和商品。

推荐系统[3]按照不同的推荐机制主要分为基于内容(content-based)的推荐、基于协作过滤(collaboration filtering)的推荐、基于知识(knowledge-based)的推荐和混合推荐等。其中基于协作过滤的算法应用得最为广泛,它的主要思想是利用已有用户群体过去的行为或者意见预测当前用户最可能喜欢哪些东西或者对哪些东西感兴趣。协作过滤推荐算法分为基于用户(User-based)的推荐算法和基于项目(Item-based)的推荐算法。这两种算法都是基于用户-项目评分矩阵寻找相似用户或者相似项目来给出推荐列表。在实际的推荐过程中,需要解决冷启动和数据稀疏性带来的推荐精度不高的问题。

处理稀疏性问题可以利用用户的特征属性,比如年龄、性别、职业等[2]。用户特征在一定程度上反映了用户的偏好,在协作过滤算法中引用用户特征可以有效解决冷启动问题;同时引入专家信任协作过滤,使得预测结果在相似背景用户和专家意见之间取得平衡。

本文第2节讨论了相关研究;第3节介绍了研究方法;第4节提出了改进的算法;第5节给出了实验结果及分析;最后总结全文并对未来工作进行展望。

2 相关研究

为了解决数据稀疏性问题,提高预测推荐的准确性,国内外大量学者提出了不同的改进方法。文献[4]综合不同用户特征和项目属性为用户生成推荐,提高了推荐算法的精确度。文献[5]针对最近邻参数 k 优化问题,提出了通过粒子群算法对参数进行在线优化和选择。文献[6]同时考虑用户特征和用户不同时间的兴趣,提出了一种基于用户特征和时间的协作过滤算法,使得越接近采集时间的用户兴趣拥有越大的权重。文献[7]提出了新的计算相似度的方法,该方法根据不同目标项目选择不同的邻居。

到稿日期:2015-10-10 返修日期:2016-01-12 本文受国家自然科学基金项目(61462022)资助。
高发展(1990—),男,硕士生,主要研究方向为个性化推荐、数据挖掘,E-mail:1304379554@qq.com;黄梦醒(1973—),男,教授,博士生导师,主要研究方向为数据与知识工程、云计算与物联网、个性化服务等;张婷婷(1991—),女,硕士生,主要研究方向为个性化推荐。

上述文献主要通过用户之间的共同评分项目集得到相似性计算方法。然而在数据极端稀疏的情况下，用户共同评分的项目集可能较小，会造成相似性计算的准确率较低的问题。用户相似度还可以考虑将用户特征和专家信任加入到计算中。文献[8]首次将专家信任引入到相似性计算中，通过计算用户和专家集的相似度，为用户预测推荐项目集。实验表明，在相同的精确度情况下，将专家信任加入到协作过滤中能有更好的用户体验。文献[9]利用专家意见提出了一种新的相似度计算算法，该算法结合了用户的评分和专家意见，并在专家意见和相似用户之间预测一个正确的平衡，从而提高推荐精度。但上述文献还存在一些不足之处：1)在数据集稀疏的情况下，用户仅仅通过用户-项目评分矩阵很难准确地预测出未评分的项目评分值；2)在考虑专家信任的同时，仅仅利用用户与专家数据集计算相似性，而没有考虑到不同特征的用户所在群体的普遍评分，因为专家和用户的背景信息不同，所以他们对同一事物的意见也会不一样。以评价一部电影为例，专家衡量一部电影的好坏是通过专业的标准来评价的，而普通用户对同样一部电影，可能会因为年龄、职业、爱好的不同而做出不同的评价，这就造成了专家意见和用户意见不统一，降低了推荐质量。本文受此启发，不仅考虑到用户特征，还考虑到专家信任，通过分析用户的不同特征，进一步比较用户与专家的相似度，从而降低数据集的稀疏性，提高预测的准确性。

3 研究方法

3.1 用户-项目评分矩阵

用户对项目的评分数据可以用一个 $m*n$ 的矩阵表示，如表 1 所列，其中行表示 n 位用户，列表示 m 条项目，r_{ij} 表示用户 i 对项目 j 的评分，$1\leqslant i\leqslant m, 1\leqslant j\leqslant n$，$r_{ij}$ 的值在 0～5 分之间。若用户 i 对项目 j 没有评分，则 $r_{ij}=0$。

表 1 用户-项目评分矩阵

用户	项目 m				
	Item_1	…	Item_j	…	Item_m
User_1	r_{11}	…	r_{1j}	…	r_{1m}
…	…	…	…	…	…
User_i	r_{i1}	…	r_{ij}	…	r_{im}
…	…	…	…	…	…
User_n	r_{n1}	…	r_{nj}	…	r_{nm}

3.2 相似度计算

基于用户的协作过滤算法中，用户相似度之间的度量方法[3]主要有 3 种：余弦相似性、修正余弦相似性和相关相似性。本文选用相关相似性作为用户相似性的度量方法。

相关相似性又称为 Pearson 相关相似性，其度量方法用式(1)表示：

$$sim(u,v)=\frac{\sum_{i\in I_{uv}}(r_{ui}-\overline{r_u})(r_{vi}-\overline{r_v})}{\sqrt{\sum_{i\in I_{uv}}(r_{ui}-\overline{r_u})^2}\sqrt{\sum_{i\in I_{uv}}(r_{vi}-\overline{r_v})^2}} \tag{1}$$

其中，$\overline{r_u}$和$\overline{r_v}$分别表示用户 u 和用户 v 在共同项目 I_{uv} 上的平均评分。

3.3 基于用户特征的协作过滤

传统基于用户的协作过滤算法仅仅利用用户对项目的评分来计算用户相似度，而没有考虑到用户的不同特征存在的差异，如用户年龄、性别等。由于不同特征的用户拥有的兴趣爱好可能不同，而同一类别的用户的兴趣爱好具有一定的相似性[2]，因此将用户特征加入到传统的相似度计算中可以提高推荐精度，同时也能解决一定的冷启动问题。研究表明，当数据集中几乎没有用户评分数据时，利用用户特征也能启动协作过滤算法。所以当新用户到达推荐系统时，利用用户填写的注册信息可以有效缓解推荐系统中的冷启动问题。用户特征用特征向量表示为$(a_1,a_2,a_3,\cdots,a_n)$。

用户性别：大多数情况下，不同性别的用户的会选择不同类别的项目。比如，女性用户偏向于情感片，男性用户偏向于武打动作片。所以将性别分为男性、女性。

0：代表男性；

1：代表女性。

用户年龄：不同年龄段的用户需求也有所不同，比如，小孩喜欢看动画片，青少年喜欢看校园片，中年人喜欢看家庭生活片，老年人喜欢看纪录片等。通过统计分析不同年龄段的喜好差异，给出了 7 个年龄段的划分。

1：0—6 岁；

2：7—12 岁；

3：13—18 岁；

4：19—30 岁；

5：31—40 岁；

6：41—60 岁；

7：60 岁以上。

用户属性的相似度计算方法如式(2)所示：

$$dem\_cor(u,v)=\sum_{k=1}^{n}sim(u_k,v_k)\times w(a_k);\ \sum_{k=1}^{n}w(a_k)=1 \tag{2}$$

其中，n 是用户属性的维度；$sim(u_k,v_k)$是用户 u 和用户 v 在属性 a_k 上的相似度；$w(a_k)$是属性 a_k 所占的权重，该权重一般由实际统计数据产生。用户 u 和 v 的属性相似度 $dem_cor(u,v)$等于在所有属性上的相似度的加权和。如果 a_k 是数值属性，不同属性上的相似度按照式(3)进行计算：

$$sim(u_k,v_k)=\frac{1}{|u_k-v_k|+1} \tag{3}$$

3.4 基于专家信任的协作过滤

2009 年 Xavier Amatriain 等人提出了基于专家信任的协作过滤算法[8]，该算法中只需计算活动用户与少量专家用户之间的相似度即可，可以大大降低计算的复杂度。基于专家信任的协作过滤算法，需要从系统中选出一批专家用户，计算普通用户与专家的相似度，建立相似度矩阵，并从专家集中选出 k 个专家作为普通用户的相似近邻，然后根据相似近邻为普通用户做出预测、推荐。

计算用户 u 和专家 e 之间的相似度 $sim(u,e)$的公式如下：

$$sim(u,e)=\frac{\sum_{s\in S_{ue}}(r_{u,s}-\overline{r_u})(r_{e,s}-\overline{r_u})}{\sqrt{\sum_{s\in S_{us}}(r_{u,s}-\overline{r_u})^2\times\sum_{s\in S_{ue}}(r_{e,s}-\overline{r_u})^2}}\times\frac{2|I_{ue}|}{|I_u|+|I_e|} \tag{4}$$

其中，$|I_u|$和$|I_e|$分别表示用户 u 与专家 e 的评论项目集合。

通过选择用户 u 最近邻的专家集，然后选取与用户 u 相似度最邻近的 k 位专家形成相似的近邻集 $S(u)$，这样用户 u 对项目 i 的预测评分值 $r_{u,i}$为：

$$r_{u,i}=\overline{r_u}+\frac{\sum_{e\in S(u)}[sim(u,e)\times(r_{e,i}-\overline{r_e})]}{\sum_{u\in S(u)}|sim(u,e)|} \tag{5}$$

其中,$\overline{r_u}$和$\overline{r_e}$分别表示某个用户 u 与专家 e 的评分平均值,$r_{e,i}$ 表示专家 e 对项目 i 的评分。

4　基于改进的用户特征及专家信任协作过滤算法

为了寻找相似背景的用户,将用户特征信息加入到相似度计算中,比如,使用 MovieLens 数据集中的用户年龄、性别、职业。文献[10]提出将两种用户相似度计算方法相结合,如式(6)所示:

$$enh\_cor(u,v)=sim(u,v)+sim(u,v)\times dem\_cor(u,v) \tag{6}$$

其中,$enh_cor(u,v)$表示用户 u 和用户 v 的总体相似度,$sim(u,v)$表示用户 u 和用户 v 的评分相似度,$dem_cor(u,v)$表示用户 u 和用户 v 的用户特征相似度。

通过使用用户和专家数据集,可以得到用户 u 的两个最近邻居集,即专家最近邻居集($S\_E(u)$)和用户最近邻居集($S\_U(u)$)。用户 u 对项目 i 的预测评分如式(7)所示:

$$P_{u,j}=\overline{r_u}+\alpha\times\frac{\sum_{e\in S\_E(u)}[sim(u,e)\times(r_{e,i}-\overline{r_e})]}{\sum_{e\in S\_E(u)}|sim(u,e)|}+(1-\alpha)\times\frac{\sum_{v\in S\_U(u)}[enh\_cor(u,v)\times(r_{v,i}-\overline{r_v})]}{\sum_{v\in S\_U(u)}enh\_cor(u,v)} \tag{7}$$

其中,$\overline{r_u}$,$\overline{r_v}$,$\overline{r_e}$分别表示用户 u、用户 v 与专家 e 的平均评分值;$r_{v,i}$表示用户 v 对项目 i 的评分;α 是设定的阈值,表示影响专家意见和用户评分的比列调节因子。

本文提出了混合推荐算法,即基于改进的用户特征及专家信任的协作过滤算法。该算法改进了协作过滤算法的冷启动和数据稀疏性问题,具体步骤如下:

1)建立相似度矩阵,计算用户之间的相似度,并按照式(4)将用户特征加入相似度计算中;

2)计算用户和专家之间的相似度;

3)按照相似度从大到小排序,分别找到基于用户和基于专家的最相近的 n 个用户和专家作为目标用户的邻居;

4)根据式(7),基于最近邻居预测目标用户对待推荐项目的评分;

5)按照项目评分对项目排序,得到最终的 Top-N 推荐。

5　实验结果及分析

5.1　数据集和度量标准

采用 MovieLens 数据集(http://www.grouplens.org)对算法的有效性进行实验。MovieLens 是由明尼苏达州大学的 GroupLens 项目组开发的,该数据集包含了 943 名用户对 1682 部影片的评价,每位用户至少为 20 部电影做过评价,评分值为(1,2,3,4,5)中的任意值,评分值越大,表示用户对这部电影越喜欢[5]。用户-项目的评分矩阵的稀疏性为:

$100000/(943\times1682)\times100\%=6.3\%$

专家数据集来自 Rotten Tomatoes。根据 MovieLens 数据集中电影的标题,得到了 154 名专家对 1205 部影片的 26739 次评价。专家-项目的评分矩阵稀疏性为:

$26739/(154\times1205)\times100\%=14.41\%$

本文采用平均绝对偏差[11](MAE)作为度量标准。该标准是通过比较预测值与用户实际的评分值之间的偏差来衡量预测结果的准确性。MAE 越小,表明推荐质量越高。设预计的用户评分集合表示为$\{p_1,p_2,\cdots,p_n\}$,对应的实际用户评分集合为$\{q_1,q_2,\cdots,q_n\}$,则平均绝对偏差 MAE 的定义如式(8)所示:

$$MAE=\frac{\sum_{i=1}^{n}|p_i-q_i|}{n} \tag{8}$$

5.2　实验设计及结果分析

本实验的目的是将综合用户特征及专家信任的协作过滤算法与传统的协作过滤算法及专家信任算法进行比较。选择传统的基于用户的协作过滤算法,并与文献[8]提出的算法进行比较分析。为了提高仿真结果的真实性,实验从 MovieLens 中随机抽取 300 名用户和专家组成实验数据集,记为 DB300,并将数据集按照 8:2 的比例分为训练集和测试集。为了找到合适的近邻 k,实验将分别选取 10,20,30,40,50,60,70 和 80 作为近邻数值。

(1)合适的近邻 k 和 α

实验将用 Pearson 相关相似性作为相似性度量方法,并且将预测评分调节因子的阈值 α 分别设定为 0,0.1,…,1。当分别选择 k 为 10,20,30,50,70 时,可得到如图 1 所示的预测结果。

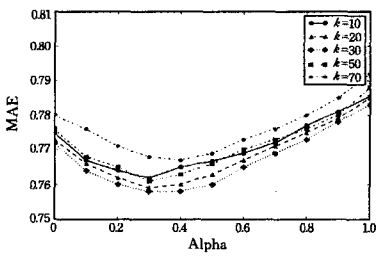

图 1　近邻值 k 和阈值 α

图 1 的实验结果反映了近邻值 k 和阈值 α 以及平均绝对误差 MAE 之间的关系。从实验结果数据的变化趋势来看,MAE 随着阈值 α 的增加呈现先下降后上升的趋势。从结果看出,当近邻数 $k=30$,$\alpha=0.3$ 时,平均绝对误差 MAE 最低,能进行更好的推荐。

(2)各种算法的比较

图 2 示出了各种协作过滤算法的比较。

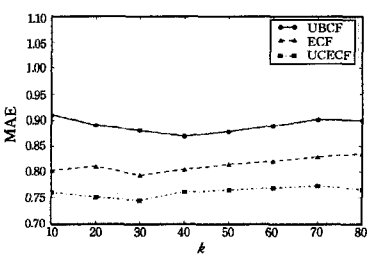

图 2　各种协作过滤算法的比较

图 2 中的横轴代表近邻数 k,纵轴代表平均绝对值 MAE。根据图 1 中的结果,将本文算法的调节因子 α 阈值设定为 0.3,并且将本文算法(UCECF)和传统的基于用户的协作过滤(UBCF)、基于专家信任的协作过滤(ECF)算法进行对比,实验结果如图 2 所示。可以看出,相对于基于用户的协作过滤算法和基于专家信任的协作过滤,本文算法具有较低的

MAE,从而预测精度更高。本文算法既考虑到用户特征对预测评分的影响,又考虑到专家信任对预测评分的影响,并且在近邻 $k=30$ 时,能够获得最佳的推荐效果。

结束语 本文针对传统协作过滤算法面临的稀疏性问题提出了综合用户特征及专家信任的协作过滤算法。当新用户到达推荐系统时,利用用户填写的注册信息可以有效缓解推荐系统中的冷启动问题。同时该算法根据不同用户特征有不同的兴趣偏好,并结合专家意见,利用协作过滤算法进行预测推荐。在 MovieLens 数据集上的实验验证了本文提出的综合用户特征及专家信任的协作过滤算法的有效性。该算法提高了推荐精度,同时缓解了数据集的稀疏性和冷启动等问题。在今后的学习工作中,可以针对不同数据集进行分析,以得到最优参数值,从而进行更好的推荐。

参 考 文 献

[1] SHAFER,SEN S W,FRANKOWSKI,et al. Collaborative Filtering Recommender Systems[C]// International Conference on Intelligent Systems Design & Applications. IEEE,2015:438-443.

[2] YANG C,AI C C,JIANG B,et al. Demographic Attribute-based Collaborative Filtering Algorithm[J]. Journal of Chinese Computer Systems,2015,36(4):782-786. (in Chinese)
杨超,艾聪聪,蒋斌,等. 一种融合人口统计属性的协同过滤算法[J]. 小型微型计算机系统,2015,36(4):782-786.

[3] JANNACH D,ZANKER M,FELFERNIG A,et al. Recommender Systems:An Introduction[M]. Int. j. hum. comput. interaction,2010.

[4] SUN L F,HUANG M X. Collaborative filtering recommendation algorithm based on user characteristics and item attributes [J]. Application Research of Computers,2014,31(2):384-387. (in Chinese)
孙龙菲,黄梦醒. 综合用户特征和项目属性的协作过滤推荐算法[J]. 计算机应用研究,2014,31(2):384-387.

[5] SHUO L X,CHAI B F,ZHANG X D. Collaborative filtering algorithm based on improved nearest neighbors[J]. Computer Engineering & Applications,2015,51(5):137-141. (in Chinese)
硕良勋,柴变芳,张新东. 基于改进最近邻的协同过滤推荐算法[J]. 计算机工程与应用,2015,51(5):137-141.

[6] PENG D W,HU B. A Collaborative Filtering Recommendation Based on User Characteristics and Time Weight[J]. Journal of Wuhan University of Technology,2009,31(3):24-28. (in Chinese)
彭德巍,胡斌. 一种基于用户特征和时间的协同过滤算法[J]. 武汉理工大学学报,2009,31(3):24-28.

[7] CHOI K,SUH Y. A new similarity function for selecting neighbors for each target item in collaborative filtering[J]. Knowledge-Based Systems,2013,37(1):146-153.

[8] 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]//Proceedings of International ACM SIGIR Conference on Research & Development in Information Retrieval. 2009:532-539.

[9] YUN L,YANG Y,WANG J,et al. Improving rating estimation in recommender using demographic data and expert opinions[C]// 2011 IEEE 2nd International Conference on Software Engineering and Service Science (ICSESS). IEEE,2011:120-123.

[10] BOζAλήÇ E,VOZALIS E,MARGARITIS K. Collaborative Filtering enhanced by Demographic Correlation[J]. Proceedings of the Aiai Symposium on Professional Practice in Ai Part of World Computer Congress,2004:293-402.

[11] LIU J G,ZHOU T,GUO Q,et al. Overview of the Evaluated Algorithms for the Personal Recommendation Systems[J]. Complex Systems & Complexity Science,2009,6(3):1-10. (in Chinese)
刘建国,周涛,郭强,等. 个性化推荐系统评价方法综述[J]. 复杂系统与复杂性科学,2009,6(3):1-10.

(上接第 102 页)

[16] LIU C C,DAI D Q,YAN H. The theoretic framework of local weighted approximation for microarray missing value estimation [J]. Pattern Recognition,2010,43(8):2993-3002.

[17] RAGEL A,CRÉMILLEUX B. MVC—a preprocessing method to deal with missing values[J]. Knowledge-Based Systems,1999,12(5/6):285-291.

[18] AGRAWAL R,SRIKANT R. Mining Quantitative Association Rules in Large Relational Tables[C]// ACM SIGMOD Conf. Management of Data. 1996:1-12.

[19] SCHNEIDER T. Analysis of Incomplete Climate Data:Estimation of Mean Values and Covariance Matrices and Imputation of Missing Values[J]. Journal of Climate,2001,14(5):853-871.

[20] RAGEL A,CREMILLEUX B. Treatment of Missing Values for Association Rules[M]// Research and Development in Knowledge Discovery and Data Mining. Springer Berlin Heidelberg,1998:258-270.

[21] GUSTAVO E,BATISTA A P A,Monard M C. An Analysis of Four Missing Data Treatment Methods for Supervised Learning [J]. Applied Artificial Intelligence,2003,17(5):519-533.

[22] RAHMAN M G,ISLAM M Z. Missing value imputation using decision trees and decision forests by splitting and merging records:Two novel techniques[J]. Knowledge-Based Systems,2013,53(9):51-65.

[23] RUBIN D B. Inference and missing data[J]. Biometrika,1976,63:581-592.

[24] LISTING J,SCHLITTGEN R. A Nonparametric Test for Random Dropouts[J]. Biometrical Journal,2003,45(1):113-127.

[25] PREISSER J S,WAGENKNECHT L E. Analysis of Smoking Trends with Incomplete Longitudinal Binary Responses[J]. Journal of the American Statistical Association,2000,95(452):1021-1031.

[26] WANG P,AN C,WANG L. An improved algorithm for Mining Association Rule in relational database[C]// 2014 International Conference on Machine Learning and Cybernetics (ICMLC). IEEE,2015:247-252.

[27] KONONENKO I,BRATKO I,ROSKAR E. Experiments in automatic learning of medical diagnostic rules[R]. Yugoslavia:Ljubljana,Lozef Institute,1984.