基于人口统计学的改进聚类模型协同过滤算法

王媛媛 李 翔

(淮阴工学院计算机与软件工程学院 淮安 223003) (河海大学计算机与信息学院 南京 211100)

摘 要 针对传统基于用户的协同过滤推荐算法在大数据环境下存在评分高维稀疏性、推荐精度低的问题,提出一种基于人口统计学数据与改进聚类模型相结合的协同过滤推荐算法,以提高推荐系统精度和泛化能力。该方法首先通过用户人口统计学数据属性,结合用户-项目评分矩阵计算各个用户间的相似度;然后对用户、项目进行分层近邻传播聚类,根据用户对项目的评分数据计算用户或项目之间的相似性,产生目标用户或项目的兴趣近邻;最后根据兴趣最近邻进行推荐。对 Epinions,MovieLents 等数据集进行仿真实验,仿真的结果表明,与传统的协同过滤算法相比,提出的算法提高了推荐精度,为传统的协同过滤推荐算法提供了参考。

关键词 协同过滤,人口统计学,聚类,推荐系统

中图法分类号 TP183 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.03.016

Study on Improved Clustering Collaborative Filtering Algorithm Based on Demography

WANG Yuan-yuan LI Xiang

(Faculty of Computer and Software Engineering, Huaiyin Institute of Technology, Huai'an 223003, China)

(College of Computer and Information, Hohai University, Nanjing 211100, China)

Abstract The traditional user based collaborative filtering recommendation algorithm in large data environment has the problem of high dimensional sparse and low recommendation accuracy. A collaborative filtering recommendation algorithm based on the combination of demographic data and improved clustering model was proposed to improve the accuracy and generalization ability of the recommendation system. Firstly, this method calculates the similarity among different users through the user demographic data attributes and the user-item score matrix. Secondly, hierarchical neighbor clustering of user and project, calculates the similarity between users or items by the user's score data for the project, and generates interest in a neighbor of a target user or project. Finally, according to the recent interest in the nearest neighbor to recommend. Simulation experiments on Epinions and MovieLents data set, the simulation results show that the proposed algorithm improves the recommendation accuracy compared with the traditional collaborative filtering algorithm, provide reference for the traditional collaborative filtering recommendation algorithm.

Keywords Collaborative filtering, Demography, Clustering, Recommender systems

推荐系统(Recommender Systems)是一种根据用户历史数据建立用户兴趣模型,协助用户过滤不相关信息,可为用户提供最佳的数据以满足用户个性化需求的信息系统[1]。推荐技术近几年已成为国内外研究和应用的热点,在电子商务、在线影视、新闻媒体等各领域均有广泛应用,如亚马逊购物(Amazon)、淘宝网(Taobao)、优酷视频(Youku)、搜狐新闻(Sohu)等。推荐系统可以辅助企业实现个性化营销,提升服务质量和产品销量,为企业创造最大的利润。

根据所使用的推荐算法,推荐系统主要分为以下几类:基于用户行为的推荐系统、基于项目内容的推荐系统、基于语境感知的推荐系统以及基于人口统计学的推荐系统等[1]。其中,基于用户行为推荐算法中的基于用户的协同过滤作为比较成功的推荐算法受到了最广泛的关注[2]。随着移动应用的迅速发展,信息数据量呈指数级增长[3],在大数据环境下,推荐系统一般会涉及社会网络数据、人口统计学数据、语境感知等多方面数据,这些多源数据一般为高维稀疏性数据,数据存在噪声和高冗余。研究表明:大数据环境下使用混合推荐算法的推荐准确度高于单独使用一种推荐算法的[4]。

传统推荐系统的输入数据规模、冗余度及噪声较小,数据稀疏性容易解决,使用协同过滤算法推荐效果较好;大数据环境下的数据规模更大,数据稀疏性、冗余度、噪声更强[5]。本文提出使用人口统计学方法统计聚类计算用户间相似度,解决大数据环境下的数据稀疏性问题,从而提高推荐准确度。

到稿日期:2015-10-01 返修日期:2016-02-20 本文受国家自然科学基金(61403060),江苏重点研发计划-产业前瞻与共性关键技术(BE2015127),江苏省高校自然科学研究面上项目(15KJB520004),江苏省先进制造技术重点实验室开放基金(HGAMTL-1401),江苏省科技厅产学研联合研究项目(BY2014097),淮安市科技计划项目(HAG2015060,HAG201602,HAC201601)资助。

王媛媛(1981-),女,博士生,讲师,CCF 会员,主要研究领域为机器学习、人工神经网络,E-mail:461044170@qq.com;李 翔(1980-),男,博士生,副教授,主要研究领域为机器学习。

1 相关问题

1.1 协同过滤算法

协同过滤算法由 Goldberg 等[6]于 1992 年提出，该算法主要考虑用户和项目协同过滤，根据用户对各项目的评分数据，使用算法分析用户兴趣特征，搜索与特定用户有相似兴趣的邻居用户，分析相似用户评价，生成指定用户喜好物品的推荐值。基于用户行为的传统协同过滤推荐算法过分依赖历史数据，对历史数据质量的要求较高，若缺少新用户和新项目的评分信息，则对新用户的信息推荐准确率较低。

另外在大数据环境中，用户以及项目的评分数据相对较少，导致评分矩阵具有稀疏性，因此目前大数据环境下使用传统协同过滤推荐算法推荐的准确度不理想[7]。目前，国内外学者提出了很多为克服评分矩阵稀疏性并提高推荐准确性的改进算法，主要有以下的改进组合思路[8]：混合、加权、特征组合、变换、特征扩充以及元级别等方式。例如，文献[9]提出了用户人口统计结合专家评分的协同过滤算法，但是有些数据集中不存在专家评分数据，专家与用户的背景知识可能有区别，所以专家的评分数据存在准确性的问题；文献[10]提出使用人口统计信息分析技术融合 EM 算法进行用户聚类；文献[11]利用社交网络中的好友信任关系缓解了评分数据的稀疏性；文献[12]提出使用划分聚类改进推荐算法；文献[13]提出高维无参数的分裂层次聚类技术；文献[14]提出对多次提取的大规模的样本进行聚类处理，进而确定自然簇质心的初始位置对推荐算法进行改进；文献[15]提出一种基于边缘度密度距的聚类方法。

1.2 基于人口统计学数据的推荐

基于人口统计学的推荐是根据人口统计学数据(一般包括人的年龄、性别、国籍、民族、工作、学历、出生地等)对每个用户建立一个用户剖面(User Profile)进行聚类，系统通过聚类计算用户间相似度，得到当前用户的最近用户邻集并以这些用户作为协同过滤的计算用户集，最后系统将邻集中评分较高的项目推荐给当前用户[16]。

本文使用用户人口统计学数据计算用户之间必要相关属性的相似度，再使用文献[17]提出的改进的分层近邻传播(Hierarchical Affinity Propagation，HAP)算法对用户进行聚类处理，最后将组内所有的用户的推荐结果进行聚合，按照推荐评分数据推荐给指定用户。

2 算法设计

使用传统的协同过滤算法计算用户相似度，一般不考虑与用户、项目相关的其他属性。本文使用用户的人口统计学数据属性对用户之间的相似度进行判断，再使用改进的分层近邻传播算法对用户进行层次聚类，以达到更好的推荐效果。

1)使用联合聚类预测评分矩阵中的未知项；

2)使用用户人口统计学数据属性，并结合联合聚类结果计算各个用户间的相似度；

3)根据上一步的结果对用户、项目进行分层近邻传播聚类，由用户对项目的评分数据计算用户或项目之间的相似性，产生目标用户或项目的兴趣近邻；

4)根据兴趣最近邻预测目标用户对待推荐项目进行目标推荐。

2.1 联合聚类预测未知项

使用联合聚类预测评分矩阵中的未知项，详细步骤如下：

1)初始化评分属于某类别的概率 $p(k|u,v,r_{u,v})$，满足：$\sum_k p(k|u,v,r_{u,v})=1$。

$$p(k|u,v,r_{u,v})=\frac{[p(k|u)+\alpha]\times[p(k|v)+\beta]\times[p(r_{u,v}|k)+\theta]}{\sum_{k'}[p(k'|u)+\alpha]\times[p(k'|v)+\beta]\times[p(r_{u,v}|k')+\theta]} \tag{1}$$

其中，α,β,θ 为超参数，为避免分母为 0，均一化为0.000000001，$p(k|u)$为用户属于某类别的概率，$p(k|v)$为项目属于某一类别的概率。

$$p(k|u)=\frac{\sum_{v=V(u)} p(k|u,v,r_{u,v})}{\sum_{z'}\sum_{v=V(u)} p(z'|u,v,r_{u,v})} \tag{2}$$

$$p(k|v)=\frac{\sum_{u=U(v)} p(k|u,v,r_{u,v})}{\sum_{z'}\sum_{u=U(v)} p(z'|u,v,r_{u,v})} \tag{3}$$

2)由式(2)、式(3)重新计算 $p(k|u)$和 $p(k|v)$。

3)计算评分值概率 $p_{discrete}(r_{u,v}|k)$。

$$p_{discrete}(r_{u,v}|k)=\frac{\sum p(k|u,v,r_{u,v})}{\sum_{r'}\sum p(k|u,v,r'_{u,v})} \tag{4}$$

4)选择概率最大的 k 作为此评分的类别，循环步骤 2)至收敛。

2.2 基于人口统计学数据计算用户相似度

用户间的相似性计算是目前推荐算法的关键，其准确性直接影响到推荐的准确性。传统协同过滤推荐算法计算用户相似性的主要方法有基于 Spearman 相关系数的相似度、基于夹角余弦的相似度、基于 Jaccard 相关系数的相似度、基于 Tanimoto 相关系数的相似度、修正余弦相似度以及绝对指数相似性等计算方法[10]。但是这类方法在大数据环境中的数据稀疏概率较高，本文结合人口统计数据计算相似度。

用户相关的人口统计数据可以反映用户偏好，结合此类信息计算用户相似度的准确性更高[10]。文献[19]的研究发现用户的人口统计数据属性如性别、年龄、职业、文化程度、地理位置、收入水平等特征信息对用户的兴趣偏好有影响。本文根据上述特征维度属性进行用户聚类。用户人口统计属性向量为$(d_1,d_2,\cdots,d_k,\cdots,d_n)$，先计算用户在每一维属性上的相似度，再结合需要使用的属性计算最后的相似度。文本考虑在数据稀疏情况下使用文献[19]提出的相似度计算方法：

$$sim(p,q)=\sum_k^n[sim(p_{dk},q_{dk})]\times w(d_k) \tag{5}$$

其中，n 为用户属性个数，$sim(p_{dk},q_{dk})$为用户 p 和 q 在 d_k 属性的相似度，$w(d_k)$是 d_k 属性的权值。用绝对指数相似性计算 $sim(p_{dk},q_{dk})$[20]，公式如下：

$$sim(p_{dk},q_{dk})=e^{-\sum_{m=1}^{m}|r_{m,p}-r_{m,q}|} \tag{6}$$

相对权值 $w(d_k)$是 d_k 属性区别不同用户的能力，那么用户在 d_k 属性两个维度之间评分最高的 t 个项目不相同的平均个数是$ave(d_k)$，权值 $w(d_k)$为：

$$w(d_k)=\frac{ave(d_k)}{\sum_{k=1}^{n} ave(d_k)} \tag{7}$$

结合以上 3 个公式计算任意两个用户在人口统计学数据中的用户相似度值。

2.3 HAP 用户聚类算法

HAP 聚类方法主要是分层获取数据集的聚类中心。首先在各个数据子集中分别执行 AP 聚类，得到子集的聚类中心后再对子集的聚类中心执行聚类，最终得到原数据集的聚类中心；然后以各聚类中心为初始类，将数据元素重新划分至与其相似度最大的聚类中心所在的类，最终实现聚类[21]。

对于任意用户属性 i，计算其他用户对其的吸引度 $r(i,j)$ 和归属度 $a(i,j)$。HAP 算法的核心是 $r(i,j)$ 和 $a(i,j)$ 两个值的不断更新，公式为：

$$r(i,j)=s(i,j)-\max_{k\neq j}\{a(i,k)+s(i,k)\} \tag{8}$$

$$a(i,j)=\min(0,r(j,j)+\sum_{k\neq\{i,j\}}\max(0,r(k,j))),i\neq j \tag{9}$$

2.4 基于人口统计学数据的用户聚类

在基于人口统计学数据计算用户相似度值的基础上，使用分层近邻传播聚类算法对用户进行聚类。结果显示，同类用户比异类用户之间的属性更接近。

1）输入用户集 U 与用户相似度矩阵 D。

2）根据 2.1 节中的公式计算相似度，并从相似度矩阵中求出最大相似度：

$$sim_{\max}=\max_{u,v\in U}(sim(u,v)) \tag{10}$$

其中，u 和 v 为任意用户集中的任意两个对象。

3）若任意两个用户对象 u 和 v 的 sim 值相同，则将两个用户对象划分为同类，再使用 2.3 节中的方法进行用户聚类。

执行上述步骤，直到聚类数量达到实际应用系统的要求，再进行预测结果推荐。

2.5 预测推荐

经过基于人口统计数据的相似度计算以及分层近邻传播用户聚类，系统根据式(7)预测某类用户对项目的评分并按分值排序推荐给指定用户。

$$pred(p,i)=\frac{\sum_{q\in neigh(p)}sim(p,q)r_{q,i}}{\sum_{q\in neigh(p)}sim(p,q)} \tag{11}$$

3 实验结果与分析

3.1 实验数据集

选取 Epinions，MovieLen(1M)，MovieLen(100k) 以及 MovieLen+ 4 个真实数据集进行实验。

Epinions 数据集(http://www.epinions.com)包含了在线服务网站 epinions.com 上的 49290 个用户、139783 个物品、664824 个评分以及 487181 个朋友关系数据。

MovieLens 数据集由美国 Minnesota 大学计算机科学与工程学院 GroupLens 项目组收集 MovieLens 网站(http://movielens.umn.edu/)上大量用户的电影评分得到，评分等级为 1－5，5 表示最喜欢，1 表示最不喜欢，用户通过评分的数值表达了自己的兴趣爱好，数据集下载地址：http://www.grouplens.org/node/73。本实验中选取了 MovieLens(1M)，MovieLens(100k)以及 MovieLens＋ 3 个不同规模的数据集作为实验数据，其中 MovieLens(1M)包含了 1 million ratings from 6000 users on 4000 movies；MovieLens(100k)包含了 100000 ratings from 1000 users on 1700 movies；MovieLens＋包含了 855598 ratings from 2113 users on 10197 movies。

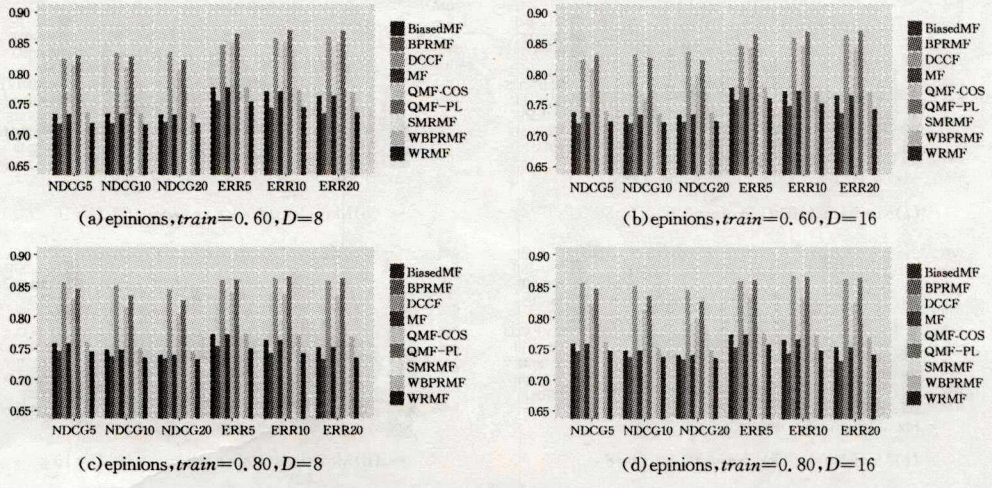
3.2 实验计算框架

本文实验采用目前流行的大数据计算框架 MapReduce，该框架可以实现对大型数据矩阵进行快速计算，也为个性化推荐系统提供计算支持。实验中在服务器上搭建 3 台虚拟机，第一台虚拟机用作 NameNode 节点，第二台虚拟机用作 SecondNameNode 节点，第三台虚拟机用作 JobTracker 节点；3 台虚拟机同时也是 DataNode 节点，模拟 Hadoop 集群的负载均衡环境。实验采用 MapReduce 和 Java 代码实现。

3.3 推荐实验结果对比

实验中选用 NDCG[22](Normalized Discounted Cumulative Gain)排名和 ERR[24-25](Expected Reciprocal Rank)作为评价指标。训练数据集随机选择 60%和 80%两种比例，项目特征维度 D 取 8 和 16 两种维度。为了比较所提出的 DCCF 方法的性能，选用 WRMF[26]，BPRMF[27]，Weighted BPRMF (WBPMF)[28]，Soft Margin Ranking MF (SMRMF)[29] 以及 Quadratic Matrix Factorization (QMF) 5 种方法做比较，同时选用 Matrix Factorization(MF)[30]，Biased Matrix Factorization (Biased MF)[31]作为基准线。

从图 1－图 4、表 1－表 4 中可以看出，本文提出的 DCCF 方法在 NDCG 和 ERR 两种评价指标中排序准确率均较高，取得了较好的结果。



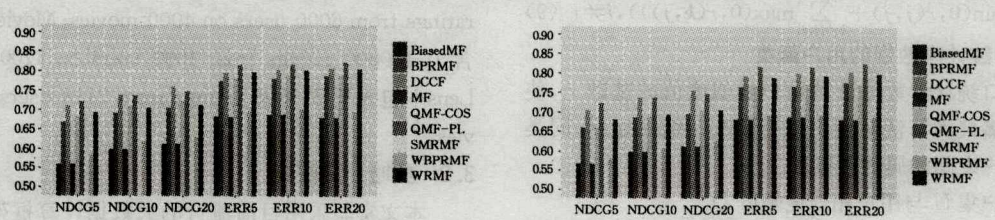
(a) epinions, $train$=0.60, D=8　　(b) epinions, $train$=0.60, D=16

(c) epinions, $train$=0.80, D=8　　(d) epinions, $train$=0.80, D=16

图 1　Epinions 数据集比较结果

表 1 Epinions 数据集比较结果

Method	*train*=0.60				*train*=0.80			
	D=8		*D*=16		*D*=8		*D*=16	
	NDCG@20	ERR@20	NDCG@20	ERR@20	NDCG@20	ERR@20	NDCG@20	ERR@20
MF	0.7344	0.7652	0.7344	0.7652	0.7393	0.7516	0.7393	0.7516
BiasedMF	0.7344	0.7652	0.7344	0.7652	0.7393	0.7516	0.7393	0.7516
WRMF	0.7212	0.7374	0.7233	0.7424	0.732	0.7343	0.735	0.7397
BPRMF	0.7216	0.736	0.7216	0.7367	0.7331	0.732	0.7326	0.7305
WBPRMF	0.7366	0.77	0.7374	0.77	0.7459	0.7676	0.7467	0.768
SMRMF	0.7239	0.7408	0.7237	0.7396	0.7339	0.7305	0.7343	0.734
QMF-PL	0.8233	0.8691	0.822	0.8692	0.8265	0.8624	0.8248	0.8634
QMF-COS	0.807	0.85	0.7996	0.8419	0.8058	0.8331	0.7994	0.8246
DCCF	0.8349	0.8604	0.8354	0.8629	0.844	0.8576	0.8433	0.8606



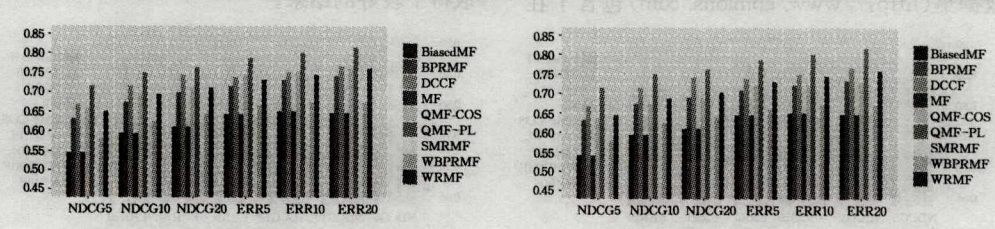
(a) Movielens(1M), *train*=0.60, *D*=8　　(b) Movielens(1M), *train*=0.60, *D*=16

(c) Movielens(1M), *train*=0.80, *D*=8　　(d) Movielens(1M), *train*=0.80, *D*=16

图 2 Movielens(1M)数据集比较结果

表 2 Movielens(1M)数据集比较结果

Method	*train*=0.60				*train*=0.80			
	D=8		*D*=16		*D*=8		*D*=16	
	NDCG@20	ERR@20	NDCG@20	ERR@20	NDCG@20	ERR@20	NDCG@20	ERR@20
MF	0.6112	0.6801	0.6112	0.6801	0.6393	0.6725	0.6393	0.6725
BiasedMF	0.6112	0.6801	0.6112	0.6801	0.6393	0.6725	0.6393	0.6725
WRMF	0.7120	0.8051	0.7049	0.7972	0.7304	0.7895	0.7265	0.7835
BPRMF	0.7043	0.7872	0.6967	0.7749	0.7304	0.7869	0.7245	0.7770
WBPRMF	0.6290	0.6956	0.6252	0.6893	0.6663	0.7010	0.6660	0.7000
SMRMF	0.6668	0.7424	0.6710	0.7524	0.7027	0.7493	0.6987	0.7427
QMF-PL	0.7470	0.8231	0.7467	0.8235	0.7710	0.8174	0.7709	0.8174
QMF-COS	0.7239	0.7885	0.7051	0.7693	0.7525	0.785	0.7335	0.7619
DCCF	0.7594	0.8059	0.7562	0.8032	0.7979	0.8134	0.7964	0.8143



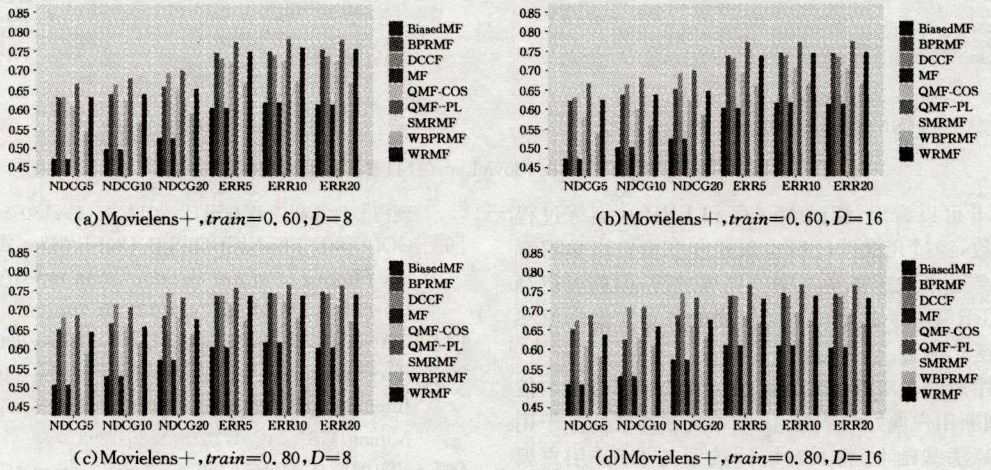
(a) Movielens(100k), *train*=0.60, *D*=8　　(b) Movielens(100k), *train*=0.60, *D*=16

(c) Movielens(100k), *train*=0.80, *D*=8　　(d) Movielens(100k), *train*=0.80, *D*=16

图 3 Movielens(100k)数据集比较结果

表 3　Movielens(100k)数据集比较结果

Method	*train*=0.60				*train*=0.80			
	D=8		*D*=16		*D*=8		*D*=16	
	NDCG@20	ERR@20	NDCG@20	ERR@20	NDCG@20	ERR@20	NDCG@20	ERR@20
MF	0.6092	0.6464	0.6092	0.6465	0.6476	0.6372	0.6476	0.6372
BiasedMF	0.6092	0.6464	0.6092	0.6465	0.6476	0.6372	0.6476	0.6372
WRMF	0.7101	0.7595	0.7026	0.7583	0.7516	0.7691	0.7448	0.7551
BPRMF	0.6961	0.7401	0.6904	0.732	0.7421	0.7561	0.7402	0.7498
WBPRMF	0.6421	0.6733	0.6388	0.6716	0.6972	0.6827	0.6911	0.6745
SMRMF	0.671	0.7239	0.6674	0.7035	0.7164	0.7057	0.7186	0.7179
QMF-PL	0.7628	0.8145	0.7619	0.8157	0.7988	0.8067	0.7992	0.8077
QMF-COS	0.7106	0.7605	0.6873	0.7287	0.7451	0.7516	0.7242	0.7237
DCCF	0.742	0.766	0.741	0.7649	0.7957	0.7692	0.7941	0.7674



(a) Movielens+, *train*=0.60, *D*=8　　(b) Movielens+, *train*=0.60, *D*=16

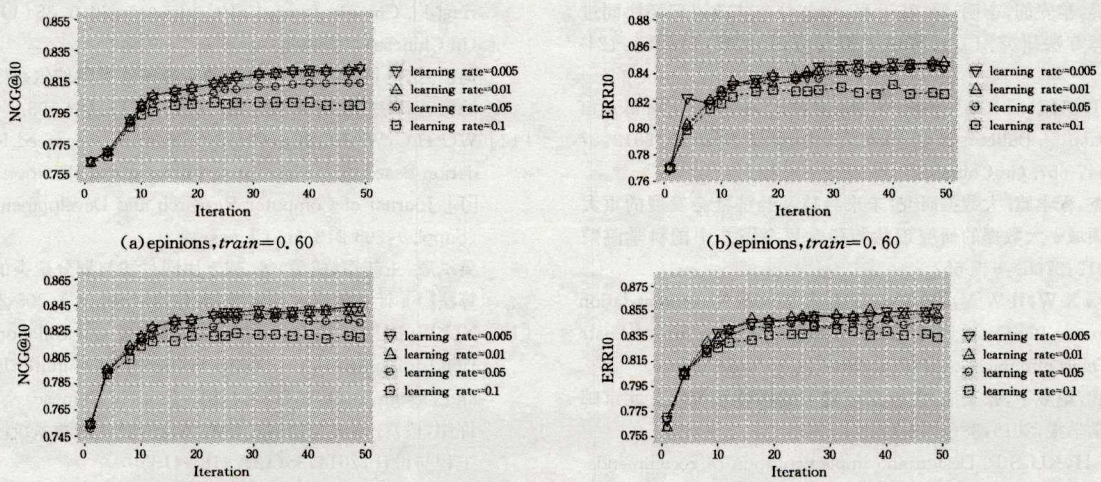
(c) Movielens+, *train*=0.80, *D*=8　　(d) Movielens+, *train*=0.80, *D*=16

图 4　Movielens+数据集比较结果

表 4　Movielens+数据集比较结果

Method	*train*=0.60				*train*=0.80			
	D=8		*D*=16		*D*=8		*D*=16	
	NDCG@20	ERR@20	NDCG@20	ERR@20	NDCG@20	ERR@20	NDCG@20	ERR@20
MF	0.5240	0.6113	0.5237	0.6127	0.5704	0.6037	0.5704	0.6037
BiasedMF	0.5240	0.6113	0.5237	0.6127	0.5704	0.6037	0.5704	0.6037
WRMF	0.6513	0.7555	0.6468	0.7460	0.6764	0.7399	0.6748	0.7304
BPRMF	0.6569	0.7530	0.6504	0.7437	0.6858	0.7469	0.6857	0.7413
WBPRMF	0.5881	0.6668	0.5839	0.6636	0.6373	0.6710	0.6314	0.6651
SMRMF	0.6092	0.6961	0.6051	0.693	0.6442	0.6847	0.6464	0.6916
QMF-PL	0.6983	0.7772	0.6989	0.7751	0.7325	0.7636	0.7328	0.7636
QMF-COS	0.6471	0.7239	0.6230	0.7034	0.6842	0.7134	0.6598	0.6909
DCCF	0.6914	0.7349	0.6911	0.7346	0.7441	0.7412	0.7426	0.7354

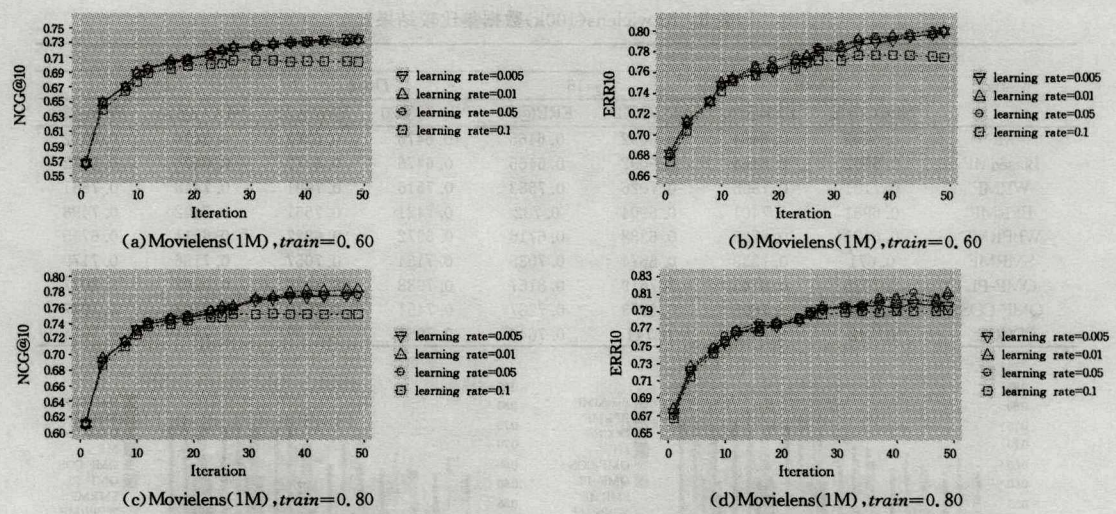
图 5 和图 6 是 NDCG@10 以及 ERR@10 在 Epinions，　MovieLens(1M)数据集上的迭代过程数据。



(a) epinions, *train*=0.60　　(b) epinions, *train*=0.60

(c) epinions, *train*=0.80　　(d) epinions, *train*=0.80

图 5　NDCG@10，ERR@10 在 Epinions 数据集上的迭代过程数据



(a) Movielens(1M), *train*=0.60 (b) Movielens(1M), *train*=0.60

(c) Movielens(1M), *train*=0.80 (d) Movielens(1M), *train*=0.80

图 6 NDCG@10, ERR@10 在 MovieLens(1M)数据集上的迭代过程数据

从图 5、图 6 可以看出，测试 NDCG 和 ERR 在训练过程的开始快速收敛，经过几次迭代后收敛速度变慢。高学习率的训练过程收敛速度快于低学习率的训练过程，但前者得到的 NDCG 和 ERR 更低。这表明本文提出的模型在低学习率下具有更好的泛化能力。

结束语 用户的人口统计数据反映了用户的部分基本情况，可以作为判断用户偏好的依据，因此本文在传统的基于用户的协同过滤算法基础上，将人口统计数据与 HAP 用户聚类推荐算法相结合，提出推荐效果更优的方法。实验分析表明，与传统的协同过滤算法相比，本文方法误差更小，有更好的推荐效果，为协同过滤推荐算法的应用研究提供了参考。

参 考 文 献

[1] ZHU Y Y, SUN J. Recommender System: Up to Now [J]. Journal of Frontiers of Computer Science and Technology, 2015, 9(5): 513-525. (in Chinese)
朱扬勇，孙婧. 推荐系统研究进展[J]. 计算机科学与探索，2015，9(5)：513-525.

[2] SUN T H, LI A N, LI M, et al. Study on distributed improved clustering collaborative filtering algorithm based on Hadoop[J]. Computer Engineering and Applications, 2015, 51(15): 124-128. (in Chinese)
孙天昊，黎安能，李明，等. 基于 Hadoop 分布式改进聚类协同过滤推荐算法研究[J]. 计算机工程与应用，2015，51(15)：124-128.

[3] LI G J, CHENG X Q. Research Status and Scientific Thinking of Big Data[J]. Bulletin of Chinese Academy of Sciences, 2012, 27(6): 647-657. (in Chinese)
李国杰，程学旗. 大数据研究：未来科技及经济社会发展的重大战略领域—大数据的研究现状与科学思考[J]. 中国科学院院刊，2012，27(6)：647-657.

[4] MENG X W, JI W Y, ZHANG Y J. A survey Recommendation Systems in Big Data[J]. Journal of Beijing University of Posts and Telecommunications, 2015, 38(2): 1-15. (in Chinese)
孟祥武，纪威宇，张玉洁. 大数据环境下的推荐系统[J]. 北京邮电大学学报，2015，38(2)：1-15.

[5] LI W H, XU S R. Design and implementation of recommendation system for E-commerce on Hadoop[J]. Computer Engineering and Design, 2014, 35(1): 130-143. (in Chinese)
李文海，许舒人. 基于 Hadoop 的电子商务推荐系统的设计与实现[J]. 计算机工程与设计，2014，35(1)：130-143.

[6] GOLDBERG D, NICHOLS D, OKI B M, et al. Using collaborative filtering to weave an information tapestry[J]. Communications of the ACM, 1992, 35(12): 61-70.

[7] TANG J, WU S, SUN J M, et al. Cross-domain collaboration recommendation[C]// Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. USA: ACM, 2012: 1285-1293.

[8] BURKE R. Hybrid recommender systems: Survey and experiments[J]. User Modeling and User-adapted Interaction, 2002, 12(4): 331-370.

[9] JIAO D J. Collaborative filtering algorithm based on user demographics and expert opinions[J]. Computer Engineering & Science, 2015, 37(1): 179-183. (in Chinese)
焦东俊. 基于用户人口统计与专家信任的协同过滤算法[J]. 计算机工程与科学，2015，37(1)：179-183.

[10] ZHANG C, CHEN G, WANG H M. Recommendation Model Based on Blending Recommendation Technology[J]. Computer Engineering, 2010, 36(22): 248-250, 253. (in Chinese)
张驰，陈刚，王慧敏. 基于混合推荐技术的推荐模型[J]. 计算机工程，2010，36(22)：248-250，253.

[11] HE J Y, MA B. Based on Real-Valued Conditional Restricted Boltzmann Machine and Social Network for Collaborative Filtering[J]. Chinese Journal of Computers, 2015, 38(1): 183-195. (in Chinese)
何洁月，马贝. 利用社交关系的实值条件受限玻尔兹曼机协同过滤推荐算法[J]. 计算机学报，2015，38(1)：183-195.

[12] WU H C, WANG X J, CHENG Y, et al. Advanced Recommendation Based on Collaborative Filtering and Partition Clustering [J]. Journal of Computer Research and Development, 2011, 48(Suppl.): 205-212. (in Chinese)
吴泓辰，王新军，成勇，等. 基于协同过滤与划分聚类的改进推荐算法[J]. 计算机研究与发展，2011，48(Suppl.)：205-212.

[13] XU W, DUAN F. Combining clustering and collaborative filtering for implicit recommender system[J]. Computer Engineering and Design, 2014, 35(12): 4181-4185. (in Chinese)
许伟，段富. 聚类与协同过滤相结合的隐式推荐系统[J]. 计算机工程与设计，2014，35(12)：4181-4185.

[14] LU Z M, FENG J G, FAN D M, et al. Novel partitional clustering algorithm for large data processing[J]. System Engineering and Electronics, 2014, 36(5): 1010-1015. (in Chinese)

卢志茂，冯进玫，范冬梅，等. 面向大数据处理的划分聚类新方法[J]. 系统工程与电子技术，2014，36(5)：1010-1015.

[15] WU M H，ZHANG H X，JIN C H，et al. Cluster Algorithm Bases on edge Density Distance[J]. Computer Science，2014，41(8)：245-249. (in Chinese)
吴明晖，张红喜，金苍宏，等. 一种基于边缘度密度距的聚类算法[J]. 计算机科学，2014，41(8)：245-249.

[16] LI G，ZHANG Z B，LIU F X，et al. Nonlinear combinatorial collaborative filtering recommendation algorithm[J]. Jouranal of Computer Applications，2011，31(11)：3063-3067.

[17] LIU X N，YIN M J，LI M T，et al. Hierarchical Affinity Propagation Clustering for Large-scale Data Set[J]. Computer Science，2014，41(3)：185-188，192. (in Chinese)
刘晓楠，尹美娟，李明涛，等. 面向大规模数据的分层近邻传播聚类算法[J]. 计算机科学，2014，41(3)：185-188，192.

[18] ALBERT R，JEONG H H，BARABÁSI A L. Attack and Error Tolerance of Complex Networks[J]. Nature，2000，406：378-382.

[19] WU Y F，WANG H R. Collaborative filtering algorithm using user background information[J]. Computer Applications，2008，28(11)：2972-2974. (in Chinese)
吴一帆，王浩然. 结合用户背景信息的协同过滤推荐算法[J]. 计算机应用，2008，28(11)：2972-2974.

[20] SUN G M，WANG S. Compute adaptive fast recommendation algorithm satisfied user interests drift[J]. Application Research of Computers，2013，30(12)：3618-3621. (in Chinese)
孙光明，王硕. 基于项目兴趣度的协同过滤新算法[J]. 计算机应用研究，2013，30(12)：3618-3621.

[21] KEPHART J，CHESS D. The Vision od Autonomic Computing[J]. IEEE Computer Society，2003，36(1)：41-50.

[22] JÄrvelin K，KekÄlÄinen J. Evaluation methods for retrieving highly relevant documents[C]// Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval(SIGIR'00). ACM，New York，NY，USA，2000：41-48.

[23] CHAPELLE O，METLZER D，ZHANG Y，et al. Expected reciprocal rank for graded relevance[C]// Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM'09). ACM，New York，NY，USA，2009：621-630.

[24] HU Y，KOREN Y，VOLINSKY C. Collaborative filtering for implicit feedback data sets[C]// Proceedings of the 2008 Eighth IEEE International Conference on Data Mining (ICDM'08). IEEE Computer Society，Washington，DC，USA，2008：263-272.

[25] GANTNER Z，DRUMOND L，FREUDENTHALER C. Bayesian personalized ranking for non-uniformly sampled items[C]// Proceedings of Knowledge Discovery and Data Mining (KDD) Cup and Workshop. 2011.

[26] WEIMER M，KARATZOGLOU A，SMOLA A. Improving maximum margin matrix factorization[J]. Mach. Learn，2008，72(3)：263-276.

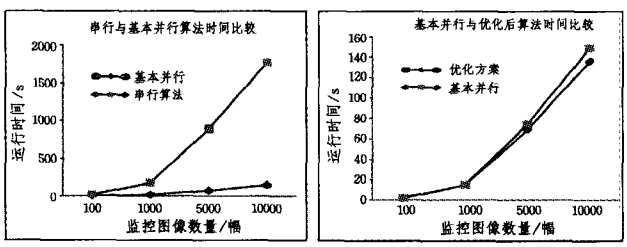
[27] RENDLE S，FREUDENTHALER C，GANTNER Z. Bayesian personalized ranking from implicit feedback [C]// Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI'09). AUAI Press，Arlington，Virginia，United States，2009：452-461.

[28] SALAKHUTDINOV R，MNIH A. Probabilistic matrix factorization[C]// Proceedings of Advances in Neural Information Processing Systems(NIPS'08). 2008：1257-1264.

[29] PATEREK A. Improving regularized singular value decomposition for collaborative filtering [C]// Proceedings of Knowledge Discovery and Data Mining (KDD) Cup and Work Shop. 2007：39-42.

[30] LIU W，WU C，Feng B，et al. Conditional preference in recommender systems [J]. Expert Syst. Appl.，2015，42(2)：774-788.

(上接第 37 页)



(a)　　(b)

图 3　并行算法时间折线图

而通过图 3(b)可以看出，优化并行算法执行时间比基本并行算法短。

结束语　经过验证，通过对算法进行分解，采取多线程处理数据的处理方式，而提取 ROI 用其余的线程并行运行的方式。在此基础上对线程进行分组，每 8 个线程一组，每组共享一个缓存队列，减少共享缓冲队列的线程数和每个缓冲区锁定的次数，以达到减少线程等待数据时间的目的。优化后的算法运行时间相比串行时间能达到大约 13.1 倍的加速。

参 考 文 献

[1] FREJLICHOWSKI D，GRZEGORZEWICZ K. An Approach to Automatic Detection and Extraction of Regions of Interest in Still Images[M]// Image Processing and Communications Challenges 4. Spring Berlin Heidelberg，2013：3-10.

[2] HIO C，BERMINGHAM L，CAI G，et al. A Hybrid Grid-based Method for Mining Arbitrary Regions-of-Interest from Trajectories[C]// The Workshop on Machine Learning for Sensory Data Analysis. 2013.

[3] EHTESHAMI N S M，TABANDEH M，FATEMIZADEH E. A new ROI extraction method for FKP images using global intensity[C]// 2012 Sixth International Symposium on Telecommunications (IST). IEEE，2012：1147-1150.

[4] SAIFULLAH A，LI J，AGRAWAL K，et al. Multi-core real-time scheduling for generalized parallel task models[J]. Real-Time Systems，2013，49(4)：217-226.

[5] LIANG H，LIU R，GUO W. Performance of the Buffer Queue With Priority For Dynamic Spectrum Access[C]// 2010 International Conference on Advanced Intelligence and Awarenss Internet (AIAI 2010). 2010：109-112.

[6] BERGAN T，CEZE L，DAN G. Input-Covering Schedules for Multithreaded Programs[J]. ACM Sigplan Notices，2013，48(10)：677-692.

[7] CHEN G，STENSTROM P. Critical lock analysis：Diagnosing critical section bottlenecks in multithreaded applications[C]// Proceedings of the 2012 International Conference for High Performance Computing，Networking，Storage and Analysis. IEEE Computer Society，2012：1-11.

[8] DICE D，MARATHE V J，SHAVIT N. Lock cohorting：a general technique for designing NUMA locks[J]. ACM Sigplan Notices，2012，47(8)：247-256.