一个解决协同过滤推荐系统相关问题的新算法

陈 琦 吕 杰 张世超
(天津大学电子信息工程学院 天津 300072)

摘 要：针对大数据应用中用户协同过滤推荐系统存在的扩展性与稀疏性问题，提出融合奇异值分解与聚类的 SBK-CF 算法。算法采用改进的皮尔逊相似度度量用户间的相似度，通过对降维后的用户进行聚类，并遍历用户的最临近簇生成推荐列表。实验结果表明，提出的算法能够有效完成个性化推荐，在一定程度上解决用户协同过滤推荐系统中存在的扩展性与稀疏性问题。

关键词：协同过滤；扩展性；稀疏性；奇异值分解；聚类
中图分类号：TN82 **文献标识码**：A **国家标准学科分类代码**：510.1050

New algorithm for collaborative filtering recommendation system-related problem solving

Chen Qi Lü Jie Zhang Shichao
(School of Electronic Information Engineering, Tianjin University, Tianjin 300072, China)

Abstract: In this paper, a fusion of singular value decomposition and clustering algorithm named SVD biKmeans collaborative filtering (SBK-CF) is proposed, in order to solve the scalability and sparsity problems in user based collaborative filtering system. The algorithm adopts improved Pearson similarity metric formula to measure similarity between users, and clustering those users which have been dimension reducted, then generates the recommendation list through the nearest neighbor cluster of users. Experimental results show that the proposed algorithm can effectively accomplish the mission of personalized recommendation and solve the scalability and sparsity problems in user-based collaborative filtering system.

Keywords: collaborative filtering; scalability; sparsity; SVD; clustering

1 引 言

随着“互联网+”的蓬勃发展和大数据时代的来临，网购已经成为人们日常生活的重要组成部分。但是快速发展的电子商务仍存在一些亟待解决的问题，“信息超载”就是一个典型的问题。由于没有充足的时间和耐心去浏览网站中所有网页，消费者通常很难准确定位到他们所满意的商品或服务，大数据时代下信息的展示与推送成为一个急需解决的问题。推荐系统能够根据消费者的个人偏好，推荐各种不同的商品或服务，从而使得消费者能够迅速且准确地找到自己满意的商品或服务。本文针对推荐系统在面对海量数据且用户评分数据不足的情况下存在的扩展性与稀疏性问题，提出了一个新的算法，算法融合了降维、改进相似度与聚类等思想，弥补传统协同过滤推荐算法的缺陷。

2 相关工作

2.1 推荐系统概述与协同过滤原理

1)推荐系统概述

什么是推荐系统，说通俗一点就是个性化定制[1]。推荐系统的基本任务是联系用户和物品，解决信息过载的问题。根据不同数据源发现数据相关性的方法推荐系统可以分为以下三种：基于人口统计学的推荐、基于内容的推荐，与协同过滤推荐。

2)协同过滤原理

协同过滤推荐的基本思想是，如果用户在过去有相同的偏好，那么他们在未来也会有相似的偏好。例如，如果用户 A 与用户 B 有着相同的购物经历，而用户 A 最近又买了一本用户 B 还不知道的书，我们就会向用户 B 推荐这本书，这种技术被称为协同过滤(collaborative filtering，CF)。

收稿日期：2015-10

2.2 协同过滤推荐系统存在的问题

1)扩展性

推荐系统中可扩展性是指问题规模可扩展性,主要关注系统在处理具有更大数据量和工作负载的更大求解问题时性能如何。大数据的分析涉及简单的统计分析以及分类汇总,其挑战在于导入数据量大,查询请求多[2],得益于信息化的深入,大数据的存在使基于过程历史数据的状态检测方法越发受到关注[3]。数据的规模越大,越需要对其进行深度挖掘,获得更多有价值的信息[4]。传统的协同过滤算法将面对较为庞大的计算量,计算的时间与空间复杂度将会越来越大;同时,较大的时间与空间开销会使得推荐时间成本加大,最终影响推荐效果。根据文献[5-7],使用聚类算法可以有效地解决推荐系统的扩展性问题并提高推荐的准确率。文献[5]提出了改进重心选择方法和距离测量的 K-均值聚类算法。文献[6]研究了基于模型的协同过滤技术的应用,特别是提出了一个集群 CF 框架和 2 个聚类算法 CF:基于项目的模糊聚类协同过滤(IFCCF)和信任感知聚类的协同过滤(TRACCF)。它们通过在 Epinions,MovieLens,Jester 和 Poste Italiane datasets 等数据集上的验证表明算法的有效性。文献[7]提出了 PCA-GAKM 算法,首先通过主成分分析(principal component analysis)预处理数据,然后通过融合遗传算法改进 K-均值聚类算法,最后应用 TOP-N 推荐机制生成推荐列表,在实验结论中验证了文章提出的算法能够有效的解决推荐系统的扩展性问题。

2)稀疏性

稀疏性问题是指在推荐系统中,由于用户和项目的数量非常巨大,相对于进行推荐预测所需的评分数量,已经获得的评分数量非常有限。稀疏性问题是协同过滤推荐算法中被广泛关注的一种经典问题,该问题一直影响传统协同过滤推荐系统的健康发展。文献[8-10]提出了一些解决推荐系统稀疏性问题的算法。其中,文献[8]提出了分类算法与相似度技术相结合的模型;文献[9]提出了可以使用用户信任的邻居评分来补充和代表用户的偏好,并发现其它相似用户;文献[10]构造了错误反映模型,应用于新的预测。

3 算法描述

3.1 算法思想

针对上面提到的扩展性与稀疏性问题,本文提出了一种结合奇异值分解与聚类的改进协同过滤推荐算法—SBK-CF 算法。

首先,本文利用奇异值分解(singular value decomposition,SVD)从 Movielens 数据集中构建一个主题空间,然后在该空间下计算相似度。SVD 将原始的评分数据矩阵 ***rdata*** 分解成 3 个矩阵 $\boldsymbol{U},\boldsymbol{V},\boldsymbol{W}$。假设 ***rdata*** 是 $m\times n$ 矩阵,那么 $\boldsymbol{U},\boldsymbol{V},\boldsymbol{W}$ 就分别是 $m\times m$,$m\times n$,$n\times n$ 矩阵,矩阵分解过程如下:$\boldsymbol{rdata}_{m\times n}=\boldsymbol{U}_{m\times m}\times\boldsymbol{V}_{m\times n}\times\boldsymbol{W}_{n\times n}$ 经过对矩阵 ***rdata*** 的分解,使其数据稀疏性得到大幅度的降低。

其次,对降维后评分矩阵 ***rdata***1 中的用户进行聚类,扫描目标用户所在簇的其他用户历史物品综合之后将评价高的物品作为推荐结果返回给目标用户。如果聚类技术与协同过滤相结合,首先将项目或用户进行聚类,将会收缩查找的空间,从而加快了利用协同过滤技术进行预测和推荐的速度[11]。在推荐系统中,使用聚类可以有效缓解扩展性问题同时提供一个准确的推荐。文献[4-7]证明了聚类协调过滤推荐算法的有效性。当为目标用户寻找相似用户时,我们不需要遍历全部数据空间,只需要从距离最近的簇中找到相邻用户,然后生成推荐集。本文使用聚类算法中的 K-means 聚类算法,K-means 算法简单灵活尤其是面对大数据的时候具有计算有效性。通常情况下,在大量的项目中,用户只会对极少数项目进行评分,从而造成大部分评分数据缺失的问题[12]。针对上面提到的问题,本文对基本 K-means 聚类算法做了如下改进。

1)使用二分 K 均值算法的结果簇的质心作为基本 K 均值的质心。

2)repeat。使用改进相似度公式将每个点指派到最近的质心,形成 K 个簇;重新计算每个簇的质心。

3)until 质心不发生改变。

改进的相似度公式如下:

$$\sin(u_x,u_y)=\frac{1}{\log(1+|N(i)|)}\frac{\sum_{h=1}^{n}(r_{u_xj_h}-\bar{r}_{u_x})(r_{u_y,i_h}-\bar{r}_{u_y})}{\sqrt{\sum_{h=1}^{n}(r_{u_xj_h}-\bar{r}_{u_x})^2}\sqrt{\sum_{h=1}^{n}(r_{u_y,i_h}-\bar{r}_{u_y})^2}} \tag{1}$$

完成以上步骤后,将会对目标用户进行 TOP-N 推荐。给定有限项集合 M,其中每个项具有 k 个属性,每个属性衡量它的一个子特征,TOP-N 推荐就是根据用户偏好函数 f,选择出 f 取值最高的 n 个项作为推荐集合[13]。遍历目标用户所在簇的其他用户的物品集,去除目标用户物品集生成推荐列表。

3.2 算法描述

算法的输入为用户-项目评分矩阵 ***rdata***、比例值 *rate*、最近邻数 K;输出为目标用户预测评分集。算法步骤如下。

1)汇总用户—评分交互数据生成评分矩阵,然后利用 SVD 降维。首先读入 Movielens 数据,进行数据预处理得到用户—评分矩阵 ***rdata***,再对 ***rdata*** 矩阵降维处理得到 ***rdata***。

2)对用户进行聚类操作,改进相似度算法。利用二分聚类初始化 16 个质心,然后利用 K-means 聚类和改进的 Pearson 相关性度量公式计算用户 u 与其他用户的相似性,最后生成 16 个簇。

3)TOP-N 协同过滤推荐。首先设置参数 $rate=0.2$,用于控制测试数据集的比例,将评分数据集随机分成训练数据集和测试数据集。然后,计算 pearson 相关系数,并求出每个测试用户的 k 个最近邻用户的索引值和相关系数。

4)完成推荐。生成推荐列表,并计算预测评分准确率指标平均绝对误差 MAE 与推荐准确率,即系统推荐的 n 个商品中用户喜欢的商品所占的比例。

4　实　　验

4.1　实验数据集与测评标准

为了验证算法的有效性,本文在 Python 编程环境下仿真验证本文提出的算法。本文实验环境为:Windows7 32 位操作系统,2 GB 内存,Intel(R) Core(TM)2 Duo CPU E7500 @ 2.93 GHz,实验程序基于 python2.7 开发。

1)实验数据集

实验数据基于 Movielens 数据集。Movielens 数据集是 GroupLens Research 采集 20 世纪 90 年代末到 21 世纪初由 Movielens 用户提供的电影评分数据,目前该系统的用户已经超过 43 000 人,用户评价的项目超 1 600 个。这些数据中包括电影评分、电影元数据以及关于用户的人口统计学数据。

2)测评标准

本文采用平均绝对误差(mean absolute difference, MAE)和推荐准确率(precision)作为测评标准。

①平均绝对误差

设测试集内目标客户的推荐数据集为 $R=\{R_{uk}k \mid k=1,2,\cdots,n\}$,目标客户的真实评价集为 $r=\{r_{uk},k \mid k=1,2,\cdots,n\}$。对于每个不为 0 的"预测-评价对"$<r\text{-}R>$,都有:

$$MAE=\frac{\sum_{u,k\in T}(R_{uk}-\bar{r}_{uk})^2}{\mid T\mid}$$

式中:T 为测试集内目标客户 A 的预测值和真实评价值都不为 0 的项目的个数。MAE 越小,推荐精度越高。预测评分的准确度指标目前有很多,这类指标的思路大都很简单,就是计算预测评分和真实评分[14]。

②推荐准确率

$$Precision=\frac{\sum_{u\in U}\mid R(u)\cap T(u)\mid}{\sum_{u\in U}\mid R(u)\mid} \tag{3}$$

式中:$R(u)$是根据用户在训练集上的行为给用户做出的推荐列表;而 $T(u)$是用户在测试集上的行为列表。推荐系统的准确率是指推荐给所有用户的物品中"相关"物品所占的比例[15]。

4.2　参数选取

1)比例(*rate*)值的选取

本文采用 Movielens1M 数据集,实验数据集被随机划分为 *trainingset* 和 *testingset* 两个部分。本文分别通过实验验证基于余弦相似性和基于 pearson 相关性的协同过滤算法在不同 *rate* 值下的 MAE 值变化趋势, *testingset* 大小对于推荐性能的影响绘制成折线图,如图 1 所示。

通过图 1 可以发现 *rate* 的值越小,平均绝对误差 MAE 就越小,推荐精度就越高。考虑到测试集过小会影响到验证的覆盖率,在数据集中取 $rate=0.2$。*trainingset* 占 80%,*testingset* 占 20%。

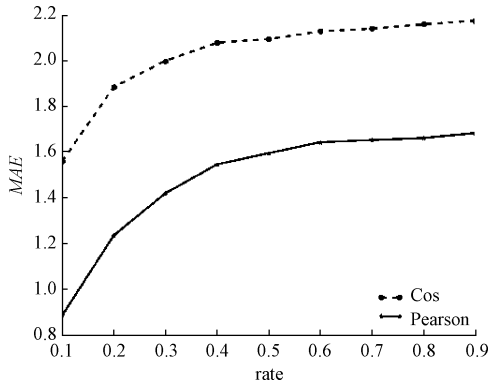

图 1　MAE 随着 *rate* 值变化图

2)相似邻居用户(K)值的选取

在 Movielens 数据集中,考虑到传统的基于用户协同过滤算法在不同 K 参数下的推荐性能表现不同。本文分别通过实验验证基于余弦相似性和基于 pearson 相关性的协同过滤算法在不同 K 值下的 MAE 值变化趋势,通过多次实验验证,K 值的大小对于推荐性能的影响绘制成折线图,如图 2 所示。

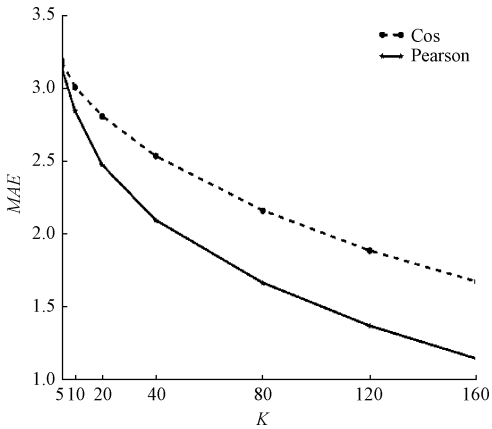

图 2　MAE 随着 K 值变化图

由图 2 可以看出,在 MovieLens 数据集中,随着 K 值的增大,平均绝对误差 MAE 的值相应的变小,推荐精度就越高。本文从推荐误差小角度考虑选择 $K=160$,即为每个用户选取的相似用户数量为 160 位。

4.3　实验结果及分析

为了证明算法的有效性,将 SBK-CF 算法与基于余弦相似度协同过滤算法和基于皮尔逊相似度协同过滤算法进行实验对比,比较结果如图 3 和 4 所示。

由图 3 可以得到,SBK-CF 算法的平均绝对误差 $MAE=0.760\ 3$,比解决扩展性与稀疏性较好的 PCA-GAKM 算法提高了 3.76%;通过图 4 我们可以得到,SBK-CF 算法的推荐准确率 $Precision=0.369\ 5$,比 PCA-GAKM

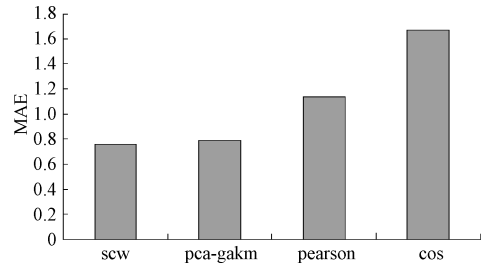

图 3　平均绝对误差的比较

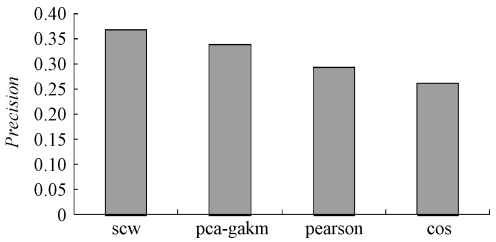

图 4　推荐准确率的比较

算法提高了 5.88%。通过对比可以看出,不管在平均绝对误差还是推荐准确率的对比上,本文提出的算法要优于 PCA-GAKM 算法。

5　结　　论

本文针对协同过滤推荐系统在互联网发展新时期所面临的扩展性与稀疏性问题,提出了一种新算法。算法采用改进的皮尔逊相似度度量用户间的相似度,通过对降维后的用户进行聚类,并遍历用户的最临近簇生成推荐列表。通过奇异值分解降维,可以提高推荐系统的效率;通过改进的协同过滤算法,能够有效地缓解扩展性与稀疏性问题。实验结果表明,本文提出的算法能够有效完成个性化推荐。

参考文献

[1] 丁大虎. 基于稀疏数据推荐系统的研究与实现[J]. 电子测量技术, 2014,37(10):119-122.

[2] 彭宇,庞景月,刘大同,等. 大数据:内涵、技术体系与展望[J]. 电子测量与仪器学报,2015,29(4):469-482.

[3] 刘吉臻,刘继伟,曾德良,等. 大数据多尺度状态检测方法在磨损检测的应用[J]. 仪器仪表学报,2013,34(1):180-186.

[4] 赵婕. 大数据时代图书馆服务探析[J]. 国外电子测量技术,2014,33(9):72-74.

[5] GHAZANFAR M A, PRüGEL-BENNETT A. Leveraging clustering approaches to solve the gray-sheep users problem in recommender systems [J]. Expert Systems with Applications, 2014, 41(7): 3261-3275.

[6] BIRTOLO C, RONCA D. Advances in clustering collaborative filtering by means of fuzzy C-means and trust[J]. Expert Systems with Applications, 2013, 40(17): 6997-7009.

[7] WANG Z, YU X, FENG N, et al. An improved collaborative movie recommendation system using computational intelligence [J]. Journal of Visual Languages & Computing, 2014, 25(6): 667-675.

[8] LIKA B, KOLOMVATSOS K, HADJIEFTHYMIADES S. Facing the cold start problem in recommender systems[J]. Expert Systems with Applications, 2014, 41(4): 2065-2073.

[9] GUO G B, ZHANG J, THALMANN D. Merging trust in collaborative filtering to alleviate data sparsity and cold start[J]. Knowledge-Based Systems, 2014, 57(2): 57-68.

[10] KIM H N, EL-SADDIK A, JO G S. Collaborative error-reflected models for cold-start recommender systems [J]. Decision Support Systems, 2011, 51(3): 519-531.

[11] 刘旭东,葛俊杰,陈德人. 一种基于聚类和协同过滤的组合推荐算法[J]. 计算机工程与科学,2010,32(12):125-133.

[12] 查九,李振博,徐桂琼. 基于组合相似度的优化协同过滤算法[J]. 计算机应用与软件,2014,31(12):324-328.

[13] 黄震华. 云环境下 top-n 推荐算法[J]. 电子学报,2015,43(1):54-61.

[14] 朱郁筱,吕琳媛. 推荐系统评价指标综述[J]. 电子科技大学学报,2012,41(2):163-175.

[15] 刘慧婷,岳可诚. 可提高多样性的基于推荐期望的 top-N 推荐方法[J]. 计算机科学,2014,41(7):270-274.

作者简介

陈琦,硕士研究生,主要研究方向为机器学习与推荐系统研究等。

E-mail:784571244@qq.com

吕杰,硕士研究生,主要研究方向为机器学习与推荐系统研究等。

张世超,硕士研究生,主要研究方向为人工智能与音频信号处理等。