Pure Mathematics 理论数学, 2020, 10(5), 404-413
Published Online May 2020 in Hans. http://www.hanspub.org/journal/pm
https://doi.org/10.12677/pm.2020.105050

Hans汉斯

An Improved Similarity Measurement Method in Collaborative Filtering Algorithm

Zijian Lian

Department of Mathematics, College of Science, Shanghai University, Shanghai
Email: 1258210173@qq.com

Received: Apr. 12th, 2020; accepted: May 2nd, 2020; published: May 11th, 2020

Abstract

In the information age, there is a huge amount of information on the Internet. While data information brings a lot of convenience to our life, it also brings the problem of information overload. Collaborative filtering (CF) algorithm emerges as a successful personalized recommendation technique and is widely used. It analyzes the behavior of users and generates recommendations by collecting the evaluation information of other users who are in line with their interests. However, the traditional recommendation algorithm has some problems such as inaccurate similarity calculation when data is sparse, cold start and scalability, which affects the application and promotion of the recommendation system. In this paper, the basic principle and implementation steps of collaborative filtering recommendation technology are studied, and an improved similarity measurement method is proposed, which can improve the accuracy of prediction by improving the utilization rate of data without complex calculation.

Keywords

Recommendation System, Collaborative Filtering, Machine Learning, K Nearest Neighbor, Similarity

协同过滤算法中一种改进相似度度量的方法

连自建

上海大学理学院数学系，上海
Email: 1258210173@qq.com

收稿日期：2020年4月12日；录用日期：2020年5月2日；发布日期：2020年5月11日

文章引用：连自建. 协同过滤算法中一种改进相似度度量的方法[J]. 理论数学, 2020, 10(5): 404-413.
DOI: 10.12677/pm.2020.105050

摘　要

信息时代，互联网上的信息量巨大，数据信息给我们的生活带来许多便利的同时，也带来了信息超载问题。协同过滤算法应运而生，作为成功的个性化推荐技术，得到了广泛的应用。它分析用户的行为，通过收集与用户兴趣一致的其他用户的评价信息来产生推荐。然而，传统的推荐算法存在数据稀疏时相似度计算不准确，以及冷启动、可扩展性问题，影响了推荐系统的应用和推广。本文研究了协同过滤推荐技术的基本原理及实现步骤，提出了一种改进的相似度度量方法，可以在不进行复杂计算的情况下，通过提高数据的使用率来很好地提高推荐的准确性。

关键词

推荐系统，协同过滤，机器学习，K近邻，相似度

1. 引言

在大数据时代，人们经常遇到信息超载的问题，搜索引擎和推荐系统是解决这一问题非常有效的工具，本文主要研究协同过滤推荐系统。如何为用户提供高质量的推荐，是推荐系统研究的重要目标。在过去十几年里，催生了非常多的推荐技术，涌现了许多的推荐系统，例如：Amazon 的个性化物品推荐、Netflix 的视频推荐、Facebook 的好友推荐，今日头条的时事新闻推荐，最近几年，抖音、快手也是国内非常火热的视频推荐系统。推荐系统是一种软件工具和技术，根据用户的兴趣特点和购买行为，为用户提供有用的项目建议，提供的建议旨在支持用户进行各种决策，广泛被大家接受的推荐系统的定义是 1997 年 Resnick 和 Varian 提出的：“它是利用电子商务网站向客户提供商品信息和建议，帮助用户决定应该购买什么产品，模拟销售人员帮助客户完成购买过程”，要向指定的用户推荐项目，系统需要收集用户偏好信息[1]。

据所获得的信息的类型，可以使用许多方法来生成建议[2] [3]。常见的推荐算法有几种：基于协同过滤的推荐系统[4] [5] [6]是一种成功的推荐系统，它利用一组相似用户的已知的和共同的偏好，对过去有相似偏好的其他用户的未知偏好做出适当的推荐。基于内容的推荐[7] [8]试图推荐与用户过去喜欢的内容相似的内容。混合推荐系统[9]利用集成技术将基于协同过滤的方法与基于内容的方法相结合。除了上述算法，还有基于知识的推荐[10]和基于效用的推荐。在这些推荐算法中，基于协同过滤的推荐算法是目前电子商务系统中使用最广泛的推荐系统。协同过滤有两种类型：基于项目的协同过滤[11]和基于用户的协同过滤[12]。这两种类型均使用识别活动用户或项目的最近邻居算法[13]。

虽然协同过滤推荐系统是一种流行的推荐系统，但它也存在一些局限性。其中一个限制是数据稀疏问题[14]，为了计算两个项目之间的相似度，它需要至少两个用户同时对相同的两个项目打分，准确的预测总是需要密集的数据，因此稀疏数据集不如密集数据集好。然而，密集数据集通常带来另一个问题，称为可扩展性问题。此外，还存在一个被称为冷启动问题的限制，它不可能为新用户或新项目找到相似的用户。当一个新项目被添加到这个系统中，没有关于它的评级信息，那么没有人可以得到关于这个项目

的推荐[15]。

协同过滤算法中的一个重要组成部分就是相似度，在不同的实现中有各种各样的相似度度量来计算两对项目之间的相似度。推荐系统中常用的相似度度量有：基于欧氏距离的相似度、余弦度量相似度、调整余弦度量相似度、皮尔逊相关相似度、Tanimoto系数相似度、Log-Likelihood等[16]。对于推荐系统来说，不同的相似度度量的选择会导致不同的结果和质量。在GroupLens和MovieLens等项目中进行了值得注意的研究。本文我们在协同过滤推荐算法中，提出了一种改进的相似度度量方法，可以在不进行复杂计算的情况下，通过提高数据的使用率来很好地提高推荐的准确性。

本文的结构如下：第 1 部分为引言。第 2 部分介绍了传统相似性度量计算方法。第 3 部分中，我们将介绍模型的实现原理与改进思想。在第 4 部分，相比于传统的相似度度量，我们评估所提出的改进度量方法。第 5 部分为结束语。

2. 常用相似度度量介绍

在本节中，我们主要描述了一些常见相似度度量。如前文所述，两个项目或两个用户之间的相似度度量是影响推荐算法效果的关键因素之一。相似度度量可以大致分为两类:基于角度的和基于距离的。我们可以用向量空间法来测量相似度度量，所有的度量都可以给出两个向量之间的相似度量的概念。考虑在n维特征空间中的P_i和Q_i两个向量，分别用和的笛卡尔坐标表示为$(P_1,P_2,\cdots,P_i,P_{i+1},\cdots,P_n)$和$(Q_1,Q_2,\cdots,Q_i,Q_{i+1},\cdots,Q_n)$。常见相似度指标的定义如下：

2.1. 基于欧氏距离的相似度

在欧式空间里，欧式距离在两个向量之间广泛使用的是笛卡尔距离，欧式距离定义如(1)式：

$$d(P_i,Q_i)=\sqrt{\sum_{i=1}^{n}(P_i-Q_i)^2} \tag{1}$$

在本文中，我们可以定义P_i向量和Q_i向量之间的相似性度量S_{PQ}^{Eu}，如下(2)式所示：

$$S_{PQ}^{Eu}=\frac{1}{1+d(P_i,Q_i)} \tag{2}$$

2.2. 基于余弦度量的相似度

在内积空间里，用P_i向量和Q_i向量之间的夹角余弦值来作为两个向量的相似性度量。由于它是两个向量之间的标准化点积，可以通过简单的数学运算来计算，因此它是一种常用的相似性度量。相应的相似性度量S_{PQ}^{Co}，计算公式如(3)式：

$$S_{PQ}^{Co}=Cosine(P_i,Q_i)=\frac{\sum_{i=1}^{n}P_i\times Q_i}{\sqrt{\sum_{i=1}^{n}P_i^2}\sqrt{\sum_{i=1}^{n}Q_i^2}} \tag{3}$$

此外，还有一种类似的修正余弦相似度S_{PQ}^{AC}定义为(4)式的形式：

$$S_{PQ}^{AC}=AdjustedCosine(P_i,Q_i)=\frac{\sum_{i=1}^{n}(P_i-\bar{S})\times(Q_i-\bar{S})}{\sqrt{\sum_{i=1}^{n}(P_i-\bar{S})^2}\sqrt{\sum_{i=1}^{n}(Q_i-\bar{S})^2}} \tag{4}$$

其中，$\overline{S}$ 是 P_i 向量和 Q_i 向量相同元素的均值。

2.3. 基于皮尔逊相关系数的相似度

皮尔逊相关系数公式是统计中最常用的公式之一，计算公式如(5)：

$$S_{PQ}^{Pe} = Pearson\left(P_i, Q_i\right) = \frac{\sum_{i=1}^{n}\left(P_i - \overline{P}\right)\times\left(Q_i - \overline{Q}\right)}{\sqrt{\sum_{i=1}^{n}\left(P_i - \overline{P}\right)^2}\sqrt{\sum_{i=1}^{n}\left(Q_i - \overline{Q}\right)^2}} \tag{5}$$

其中，$\overline{P}$ 和 $\overline{Q}$ 分别是向量 P_i 和 Q_i 的均值。

2.4. 基于 Tanimoto 系数的相似度

Tanimoto 系数，又称 Jaccard 系数，是余弦相似度的扩展，多用于计算文档数据相似度。基于它的相似度和上面的相似度都不相同。简单来说，Tanimoto 系数使用相交集与并集的比值作为相似性度量，具体计算公式如式(6)：

$$S_{PQ}^{Ta} = \frac{\left|P_i \cap Q_i\right|}{\left|P_i\right| + \left|Q_i\right| - \left|P_i \cap Q_i\right|} \tag{6}$$

除了上面提到的这些相似度度量外，还有其他一些度量，例如基于对数似然函数值相似度、基于曼哈顿距离相似度等。相似度度量在推荐系统中起着至关重要的作用，推荐算法的性能和效率往往取决于系统使用的相似度度量。

3. 模型与算法

在这一节中，我们介绍了基于物品协同过滤算法的工作原理和算法步骤，通过对第 2 部分介绍的传统相似性度量进行分析，指出了该模型的不足之处，并详细介绍了我们所提出的改进相似度度量方法是如何提高模型的计算精度的。

主要符号汇总：

$U = \{u_1, \cdots, u_a, \cdots, u_m\}$ 表示用户集

$I = \{i_1, \cdots, i_j, \cdots, i_n\}$ **表示物品集**

$K_i = \{K_{i1}, \cdots, K_{in}\}$ 表示物品 i 的相关性最高的 Top-K 个物品

$P_a = \{P_{a1}, \cdots, P_{aj}, \cdots, P_{an}\}$ 表示用户 a 对物品 I 预测的集合

$r_{a,j}$ 表示用户 a 对物品 j 的打分情况

R 表示评分项 $r_{a,j}$ 的 $m \times n$ 评分矩阵，其中 $a \in 1, \cdots, m; j \in 1, \cdots, n$

3.1. 协同过滤算法思想

推荐系统应用数据分析技术，找出用户最可能喜欢的东西推荐给用户，现在很多电子商务网站都有这个应用。目前用的比较多、比较成熟的推荐算法是协同过滤(Collaborative Filtering，简称 CF)推荐算法。同过滤推荐算法是诞生最早，并且较为著名的推荐算法，主要的功能是预测和推荐。算法通过对用户历史行为数据的挖掘发现用户的偏好，基于不同的偏好对用户进行群组划分并推荐品味相似的商品。主要分为三个部分：① 用户偏好描述，② 寻找最近邻居，③ 预测产生推荐。协同过滤推荐算法过程示意图

如图 1。

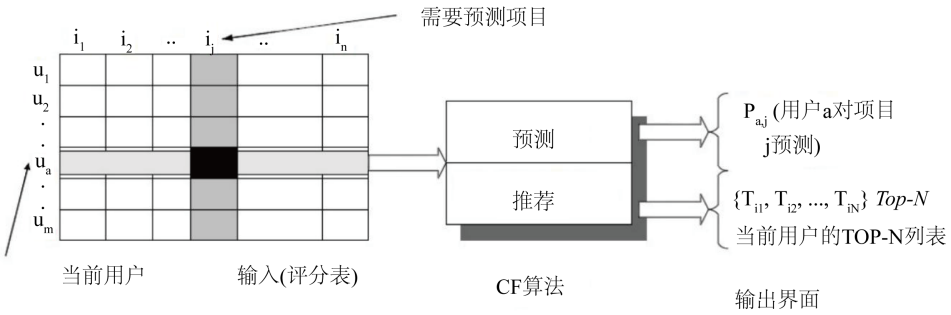

Figure 1. Collaborative filtering recommendation algorithm process
图 1. 协同过滤推荐算法过程

在 CF 中，用 $m \times n$ 的矩阵表示用户对物品的喜好情况，一般用打分表示用户对物品的喜好程度，分值定义为从 1 (非常不喜欢)到 5 (非常喜欢)，分数越高表示越喜欢这个物品，0 表示没有买过该物品。图中行表示一个用户，列表示一个物品。CF 分为两个过程，一个为预测过程，另一个为推荐过程。预测过程是预测用户对没有购买过的物品的可能打分值，推荐是根据预测阶段的结果推荐用户最可能喜欢的一个或 Top-N 个物品。

3.2. 算法步骤

算法训练版块：

输入： 训练集数据

输出： 相似度矩阵和 Top-K 近邻个物品

过程：

Step1：训练集数据预处理

Step2：建立评分矩阵

Step3：根据相似度计算评分矩阵，得到相似度矩阵和 Top-K 个物品

算法测试版块：

输入： 测试集数据

输出： 预测分数矩阵和平均绝对误差 MAE（Mean Absolute Error）

过程：

Step4：测试集数据预处理

Step5：数据预测

Step6：计算平均绝对误差

3.3. 预测模型

为了在协同过滤预测中达到预测的目的，在确定了物品间相似度后，可以对推测物品相似的 Top-K 个物品进行加权评分总和来预测物品的评分。计算预测值得方法有三种：第一种方法如式(7)所示，直接计算目标项目 K 个近邻项目的评分均值；第二种方法如式(8)所示，考虑不同近邻与目标项目之间的不同

权重的加权和，并进行标准化；第三种方法如式(9)所示，与式(8)的区别在于考虑了不同用户的评价风格问题，克服了不同用户之前的评分尺度不一样的缺点，相比而言具有更高的精确度。

$$P_{a,j}=\frac{\sum_{K_j\in K} r_{a,K_j}}{K} \tag{7}$$

$$P_{a,j}=\frac{\sum_{K_j\in K} sim\left(K_j,j\right)r_{a,K_j}}{\sum_{K_j\in K}\left|sim\left(K_j,j\right)\right|} \tag{8}$$

本文选择第三种具有上述相似性的协同过滤预测方法，形式上，我们用 $P_{a,j}$ 表示用户 a 对物品 j 的评分预测，公式(9)如下：

$$P_{a,j}=\overline{r_a}+\frac{\sum_{K_j\in K} sim\left(K_j,j\right)*\left(r_{a,K_j}-\overline{r_a}\right)}{\sum_{K_j\in K}\left|sim\left(K_j,j\right)\right|} \tag{9}$$

其中，$\overline{r_a}$ 为用户 a 的平均评分。

3.4. 改进的加权度量方法

3.4.1. 改进方法的思想原理

由于推荐系统中的数据比较稀疏，而传统的相似度度量方法只关注由两个用户同时评分的物品的评分，而不考虑他们评定的其他项目评分，致使以往的推荐算法模型对数据的使用效率不高。此外，两个用户之间的相似度相同也不一定反映真实的相关性，因为两个用户之间公同打分的物品数量不同时，也可能得到相同的相似度。

一方面，假设用户 p 和 q 对 k 个共同物品有评分，p 和 q 的相关性 $sim\left(p,q\right)=s$，而用户 u 和 v 对 j 个共同物品有评分，u 和 v 的相关性 $sim\left(u,v\right)=s$，其中 $j<k$。即便两者有相同的相关性，但直觉上感觉用户 p 和 q 之间的相似度比用户 u 和 v 之间的相似度更稳定。因此，为了更好地计算相似度指标，可以将两个用户评价的公共物品的数占自己评分总数量的比重合并到相似度评分中。

另一方面，考虑用户评价的项目总数对计算相似性度量的影响。对于任何两个用户之间，两者共同评分的项目数量一定，而用户评分项目总数越少，相似度与之越密切；相反，若用户评分项目总数越多，相似度反而也小

综上两方面，在计算相似度度量时，我们提出了一种新的方法，在计算相似度时，对用户之间共同评分数量成正比放大，对于用户评分项目总数成反比缩小，按着新的权重来计算用户之间的相似度。

3.4.2. 改进方法的实现

按着前文我们改进的加权方法思想，我们提出新的计算相似度量公式如下：

$$NewSim\left(a,b\right)=\frac{\frac{k}{t_a}Sim\left(a,b\right)+\frac{k}{t_b}Sim\left(a,b\right)}{2} \tag{10}$$

即

$$NewSim\left(a,b\right)=\left(\frac{1}{t_a}+\frac{1}{t_b}\right)\frac{k}{2}Sim\left(a,b\right) \tag{11}$$

其中，t_a 和 t_b 分别是用户 a 和 b 评分项目的总数，k 是用户 a 和 b 相同评分项目总数。

4. 实验与分析

本节中，我们对上文提出的改进相似度度量方法和传统度量方法进行评估，分别使用真实数据与传统的度量方法结果进行了分析与比较。

4.1. 数据集

为了评估我们的方法，我们使用了来自 MovieLens 数据集的实验数据。该数据集包含来自 1682 部电影的 943 名用户的 100,000 个评分(1~5)。每个用户至少为 20 部电影打分。我们随机选取 80%的数据进行训练，其余的进行测试。

4.2. 实验评估标准

推荐系统研究人员有几种方法来评估推荐算法的质量。在本文中，我们使用平均绝对误差(MAE)评估来计算实际偏好和估计偏好之间的平均差值。较低的 MAE 值意味着估计的首选项与实际首选项相差不大，MAE = 0 表示完美的预测推荐。

$$\mathrm{MAE}=\frac{\sum_{a,j}\left|r_{a,j}-P_{a,j}\right|}{n}，\text{其中 } n \text{ 为测试集总数} \tag{10}$$

4.3. 实验结果

实验结果如下表 1~4，图 2~5 所示。

Table 1. MAE before and after improvement based on Euclidean similarity
表 1. 基于欧式相似度改进前后 MAE

最近邻数量	欧式距离相似度	改进欧式相似度
25	0.8917	0.8341
50	0.8734	0.8177
75	0.8653	0.8110
100	0.8605	0.8093
125	0.8566	0.8104
150	0.8569	0.8111

Table 2. MAE before and after improvement based on Pearson similarity
表 2. 基于皮尔逊相似度改进前后 MAE

最近邻数量	皮尔逊相似度	改进皮尔逊相似度
25	0.8477	0.7826
50	0.8309	0.7767
75	0.8240	0.7702
100	0.8242	0.7646
125	0.8230	0.7615
150	0.8236	0.7613

Table 3. MAE before and after improvement based on Adjusted Cosine similarity
表 3. 基于调整余弦相似度改进前后 MAE

最近邻数量	调整余弦相似度	改进调整余弦相似度
25	0.7966	0.7502
50	0.8595	0.8036
75	0.9312	0.8644
100	1.0277	0.9424
125	1.1010	1.0050
150	1.1724	1.0545

Table 4. MAE before and after improvement based on Tanimoto similarity
表 4. 基于 Tanimoto 相似度改进前后 MAE

最近邻数量	Tanimoto	改进 Tanimoto
25	0.7423	0.7274
50	0.7526	0.7476
75	0.7621	0.7579
100	0.7688	0.7636
125	0.7733	0.7649
150	0.7751	0.7666

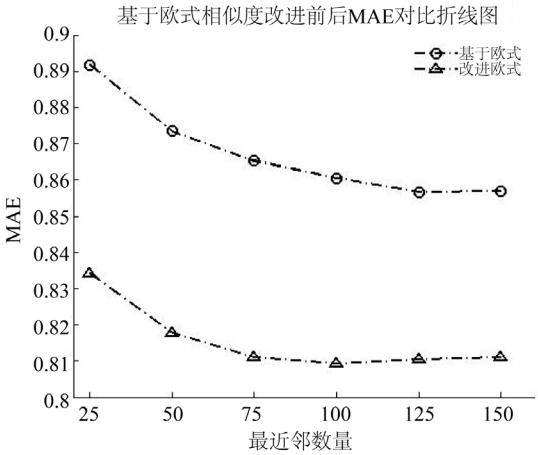

Figure 2. MAE curve before and after improvement based on Euclidean similarity
图 2. 基于欧式相似度改进前后 MAE 曲线图

4.4. 实验结果分析

由表 1~4 及图 2~5 可以得到下面结论：

1) 协同过滤推荐算法中，使用基于不同度量下的相似度会产生不同的预测精度，因此可以说明相似度度量是影响推荐算法效果的关键因素之一。

2) 协同过滤推荐算法中，在同一种相似度下，选择不同的最近邻数量也会产生不同的预测精度。

3) 基于不同度量相似度的预测精度，本文提出的改进度量方法可以在一定程度上降低各种度量下的相似度预测分数的 MAE，提高算法预测精度。

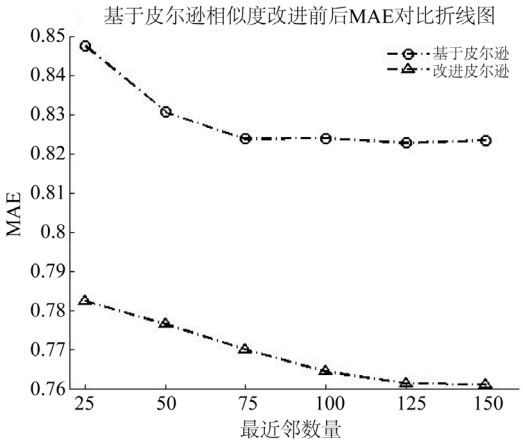

Figure 3. MAE curve before and after improvement based on Euclidean similarity
图 3. 基于皮尔逊相似度改进前后 MAE 曲线图

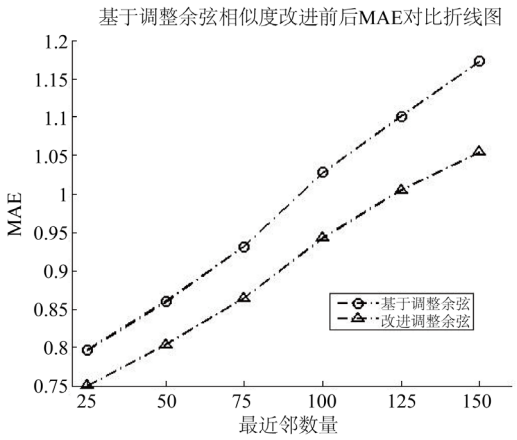

Figure 4. MAE curve before and after improvement based on Adjusted Cosine similarity
图 4. 基于调整余弦相似度改进前后 MAE 曲线图

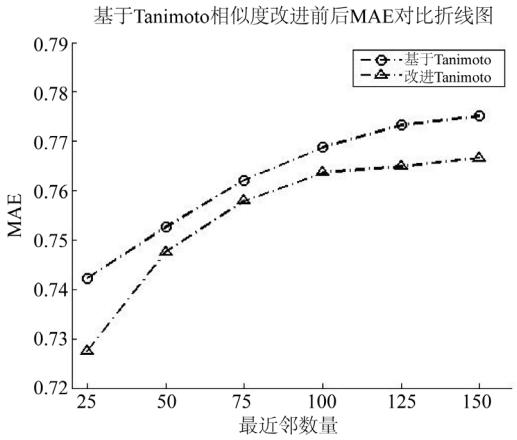

Figure 5. MAE curve before and after improvement based on Tanimoto similarity
图 5. 基于 Tanimoto 相似度改进前后 MAE 曲线图

5. 结束语

本文中，我们提出了一种改进的加权度量方法，通过提高稀疏矩阵的使用率，来提高算法中预测的精确度。在数据实验方面，我们使用推荐系统常用的数据集 MovieLens-100 K 数据集，通过与传统的几种度量方法如基于欧式距离相似度量、基于皮尔逊相关系数相似度量、基于调整余弦相似度量和基于 Tanimoto 系数相似度的预测精确度比较，我们提出的改进度量方法能有效提高传统度量的预测精度，在没有增加算法的复杂程度的情况下，普遍提升幅度在 6%~10%之间。

参考文献

[1] Ricci, F., Rokach, L. and Shapira, B. (2010) Introduction to Recommender Systems Handbook. In: Ricci, F., Rokach, L., Shapira, B. and Kantor, P., Eds., *Recommender Systems Handbook*, Springer-Verlag, New York, 1-35. https://doi.org/10.1007/978-0-387-85820-3_1

[2] Prasad, R. (2012) A Categorical Review of Recommender Systems. *International Journal of Distributed and Parallel Systems*, **3**, 73-83. https://doi.org/10.5121/ijdps.2012.3507

[3] Adomavicius, G. and Tuzhilin, A. (2005) Toward the Next Generation of Recommender Systems: Toward the Next Generation of Recommender Systems. *IEEE Transactions on Knowledge and Data Engineering*, **17**, 734-749. https://doi.org/10.1109/TKDE.2005.99

[4] 马瑞新, 孟繁成, 王涵杨. 优化的协同过滤推荐算法[J]. 计算机科学与应用, 2011, 1(3): 108-111.

[5] 崔梓凝. 基于协同过滤的推荐算法研究[J]. 数字化用户, 2017, 23(45): 160, 42.

[6] 周军锋, 汤显, 郭景峰. 一种优化的协同过滤推荐算法[J]. 计算机研究与发展, 2004, 41(10): 1842-1847.

[7] Balabanovic, M. and Shoham, Y. (1997) Fab: Content-Based, Collaborative Recommendation. *Communications of the ACM*, **40**, 66-72. https://doi.org/10.1145/245108.245124

[8] Aciar, S., Zhang, D., Simoff, S. and Debenham, J. (2007) Informed Recommender: Basing Recommendations on Consumer Product Reviews. *IEEE Intelligent Systems*, **22**, 39-47. https://doi.org/10.1109/MIS.2007.55

[9] de Campos, L.M., Fernandez-Luna, J.M., Huete, J.F. and Rueda-Morales, M.A. (2010) Combining Content-Based and Collaborative Recommendations: A Hybrid Approach Based on Bayesian Networks. *International Journal of Approximate Reasoning*, **51**, 785-799. https://doi.org/10.1016/j.ijar.2010.04.001

[10] 刘平峰, 聂规划, 陈冬林. 基于知识的电子商务智能推荐系统平台设计[J]. 计算机工程与应用, 2007(19): 203-205+220.

[11] Sarwar, B., Karypis, G., Konstan, J. and Riedl, J. (2001) Item-Based Collaborative Filtering Recommendation Algorithms. *Proceedings of the International Conference on the World Wide Web*, New York, 1-5 May 2001, 285-295. https://doi.org/10.1145/371920.372071

[12] 王成, 朱志刚, 张玉侠, 等. 基于用户的协同过滤算法的推荐效率和个性化改进[J]. 小型微型计算机系统, 2016, 37(3): 30-34.

[13] 李聪, 梁昌勇, 马丽. 基于领域最近邻的协同过滤推荐算法[J]. 计算机研究与发展, 2008, 45(9): 1532-1538.

[14] 吴颜, 沈洁, 顾天竺, 等. 协同过滤推荐系统中数据稀疏问题的解决[J]. 计算机应用研究, 2007, 24(6): 94-97.

[15] 李改, 李磊. 一种解决协同过滤系统冷启动问题的新算法[J]. 山东大学学报(工学版), 2012, 42(2): 11-17.

[16] Laveti, R.N., Ch, J., Pal, S.N. and Babu, N.S.C. (2016) A Hybrid Recommender System Using Weighted Ensemble Similarity Metrics and Digital Filters. *Processings of the* 23*rd International Conference on High Performance Computing Workshops* (*HiPCW*), Hyderabad, 2016, 32-38. https://doi.org/10.1109/HiPCW.2016.013