基于混合聚类与融合用户兴趣的协同过滤推荐算法*

麻 天[1,2],余本国[3],张 静[1,2],宋文爱[1,2],景 昱[1]
(1.中北大学 软件学院,山西 太原 030051;2.山西省军民融合软件工程技术研究中心,山西 太原 030051;
3.海南医学院 生物医学信息与工程学院,海南 海口 571199)

摘 要:推荐效率低、推荐质量有待提高等问题普遍存在于传统协同过滤推荐算法中,为了改善并解决这些问题,在协同过滤推荐算法中将混合聚类与用户兴趣偏好融合,经过验证推荐质量有显著提升。首先根据用户的个人相关信息构建 Canopy+bi-Kmeans 的一种多重混合聚类模型,采用提出的混合聚类模型把所有用户划分成多个聚类簇,将每个用户的兴趣偏好融合到生成的聚类簇中,形成新的相似度计算模型;其次利用基于 TF-IDF 算法的权重归类方法计算用户对标签的权重,并使融入时间系数的指数衰减函数捕捉用户兴趣偏好随时间的变化;最后使用加权融合将用户偏好和混合聚类模型相结合,匹配到更相似的邻居用户,计算出项目评分并进行推荐。利用公开数据集对比实验证明,提出的方法能够提高推荐质量和推荐可靠性。

关键词:推荐算法;权重标签;时间衰减系数;指数衰减函数;混合聚类
中图分类号:TP399 **文献标识码**:A DOI:10.16157/j.issn.0258-7998.212086

中文引用格式:麻天,余本国,张静,等. 基于混合聚类与融合用户兴趣的协同过滤推荐算法[J].电子技术应用,2022,48(4):29-33.
英文引用格式:Ma Tian,Yu Benguo,Zhang Jing,et al. Collaborative filtering recommendation algorithm based on hybrid clustering and user preferences fusion[J]. Application of Electronic Technique,2022,48(4):29-33.

Collaborative filtering recommendation algorithm based on hybrid clustering and user preferences fusion

Ma Tian[1,2],Yu Benguo[3],Zhang Jing[1,2],Song Wenai[1,2],Jing Yu[1]
(1.Software School,North University of China,Taiyuan 030051,China;
2.Shanxi Military and Civilian Integration Software Engineering Technology Research Center,Taiyuan 030051,China;
3.School of Biomedical Information and Engineering,Hainan Medical University,Haikou 571199,China)

Abstract:Problems such as low recommendation efficiency and recommendation quality to be improved generally exist in the traditional collaborative filtering recommendation algorithm. In order to improve and solve these problems, the collaborative filtering recommendation algorithm integrates mixed clustering with user interests and preferences, and the recommendation quality has been significantly improved after verification. Firstly, a multiple mixed clustering model of Canopy+ Bi-Kmeans was constructed according to the personal information of users. The proposed mixed clustering model was used to divide all users into multiple clusters, and the interest preferences of each user were fused into the generated clusters to form a new similarity calculation model. Secondly, the weight classification method based on TF-IDF algorithm is used to calculate the weight of users on labels, and the exponential decay function incorporating time coefficient is used to capture the change of users' interest preference with time. Finally, weighted fusion is used to combine user preferences with mixed clustering model to match more similar neighbor users, calculate project scores and make recommendations. The experimental results show that the proposed method can improve the recommendation quality and reliability.

Key words:recommendation algorithm;weight label;time attenuation coefficient;exponential decay function;hybrid clustering

0 引言

在信息快速发展的现代社会中,推荐算法已经普遍出现在人们的生活中,给人类生活无形中带来巨大便利[1],如短视频推荐[2]、音乐歌曲推荐[3]、新闻信息推荐[4]。协同过滤推荐算法在工程上更容易实现。该算法分为两类:基于用户的协同过滤推荐算法(user-based collaborative filtering)和基于项目的协同过滤推荐算法(item-based collaborative filtering)[5]。简言之:物以类聚,

*基金项目:国家自然科学基金(61602427)

人以群分。虽然协同过滤推荐算法与其他推荐算法相比有很多优点，但解决推荐效率低、推荐质量低、冷启动和稀疏矩阵等问题一直是研究者不断努力改进的方向[6]。其中在计算不同用户之间的相似性时也存在很多问题，相似度计算不精准是影响推荐准确性的一个关键因素[1]。

很多研究学者提出很多方法改进以上存在的问题。赵伟等在传统 K-means 聚类算法的基础上做了改进，有效地解决了有关用户聚类的一些问题[7]。王蓉等提出了一种混合聚类与融合属性特征的协同过滤推荐算法，在一定程度上能提高推荐效率，解决冷启动问题，为聚类算法在推荐系统中的研究开辟了新思路[6]。

本文依据上述学者的思路，改进了算法，通过建立 Canopy+bi-Kmeans 混合聚类模型[8]和一种改进的相似度计算方法，提出一种基于混合聚类与融合用户偏好的协同过滤推荐算法，从而可以达到提高推荐可靠性、提高推荐精度的效果。利用 MovieLens 数据集进行试验得出结果表明，该算法不仅能有效解决存在的冷启动问题，而且可提高推荐算法效率。

1 Canopy+bi-Kmeans 混合聚类算法

1.1 Canopy 算法

首先利用 Canopy 算法对数据集进行一次聚类，这种算法有利有弊，不需要指定 k 值，可以快速得到聚类簇，但是精度较低[9]。算法过程如下：

(1)从原始数据中生成样本列表 $X=[x_1,x_2,\cdots,x_m]$，在设定初始距离阈值 T_1、T_2 时，通过两种方式调整参数：先验知识和交叉验证，且 $T_1>T_2$。

(2)选取 Canopy 质心。从列表 X 中任选一个样本，令第一个样本为 P，并将 P 从列表中删除。

(3)从列表 X 中随机选取一个样本 R，计算 R 到所有 Canopy 质心的距离，判断其中最小的距离 D：如果 $D\leqslant T_1$，则令 R 为一个弱标记，表示 R 属于该质心，并将 R 加入其中；如果 $D\leqslant T_2$，则将 R 进行强标记，表示 R 属于该质心，更新强样本标记质心，并将样本 R 从列表 X 中移除[10]；如果 $D>T_1$，则 R 形成一个新的聚簇，并将 R 从列表 X 中删除。

(4)若列表 X 中元素个数不为零，则不断重复上述步骤(3)。

1.2 bi-Kmeans 算法

bi-Kmeans(bisecting K-means)聚类算法受随机选择初始质心的影响比较小，改进 K-means 算法随机选择初始质心的随机性造成聚类结果不确定性的问题。简言之："高内聚，低耦合"。意思是让每个类簇之间要有明显的界限，类簇内部的点要团结紧凑[11]。bi-Kmeans 算法步骤如下：

(1)从原始样本集合中随机取 k 个初始中心点。

(2)以这 k 个中心点为标准，计算所有样本点到中心的距离，计算后将其加入到距离最近的类簇。这样每个样本都有自己的簇了。

(3)重新计算每个簇中的样本中心点，如果中心点未发生变化转到步骤(4)，发生变化回到步骤(2)。

(4)得出结果。

输出：划分出的聚类簇以及聚类中心。

在选择聚类时，利用 SSE(Sum of Squared Error)当作度量聚类效果的指标。不同聚类算法对比见表 1。

表 1 不同聚类算法对比

序号	K-means	K-means++	bi-Kmeans
1	2 112	120	106
2	388	125	106
3	824	127	106
平均	1 108	124	106

从表 1 以直观地发现，bi-Kmeans 计算出来的 SSE 值最小，并且趋于稳定值，说明聚类的效果也最好。因此，本文选用 bi-Kmeans 这个聚类方法。

1.3 Canopy+bi-Kmeans 算法

Canopy+bi-Kmeans 这个聚类组合有很多优点，如增强了单独聚类抗干扰的能力，加快了相似性计算的速率。Canopy+bi-Kmeans 算法流程图如图 1 所示。

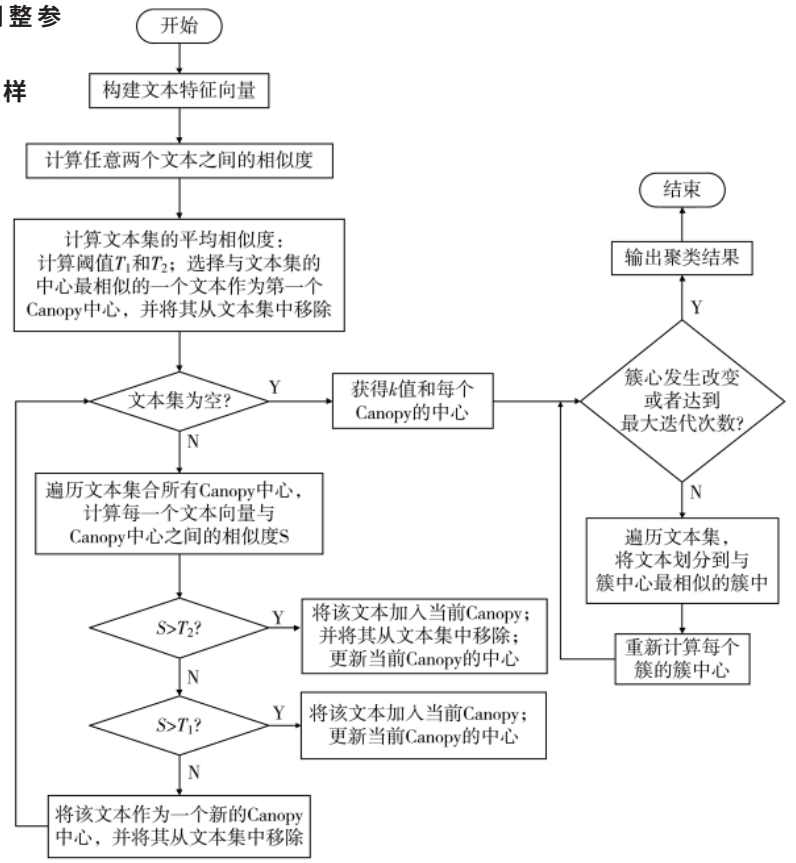

图 1 Canopy+bi-Kmeans 算法流程图

2 计算用户偏好相似性

2.1 计算用户偏好

通常用户会根据个人的兴趣对项目打分。文献[12]简单地根据标签的数量来判断用户的偏好，从而使得当前潮流标签权重过高使得某些用户选择冷门标签时无法得到更准确的推荐，未能将用户的兴趣偏好充分展现出来。这对上述问题，本文利用 TF-IDF 的方法对用户偏好进行计算。

TF-IDF 用计量统计的方式来评估某个关键词在其所在的语料库中的重要性[13]，公式如下：

$$P_{ua}=\frac{\sum_{i=1}^{n} r'_{ui}\times f_{ia}}{\sum_{i=1}^{n}\sum_{a=1}^{s} r'_{ui}\times f_{ia}}\times\lg\left(\frac{\mathrm{num}_m}{\mathrm{num}_{ua}}\times\frac{\sum_{i=1}^{n}\sum_{a=1}^{s} f_{ia}}{\sum_{i=1}^{n} f_{ia}}\right) \tag{1}$$

其中，P_{ua} 表示用户 u 对项目标签 a 的偏好值，P_{ua} 值与偏好程度成正比；n 表示项目总数，s 表示项目标签总数；$\sum_{i=1}^{n} r'_{ui}\times f_{ia}$ 表示用户 u 标注标签 a 的次数，$\sum_{i=1}^{n}\sum_{a=1}^{s} r'_{ui}\times f_{ia}$ 表示用户 u 标注的总次数；num_m 表示用户总数，num_{ua} 表示标注过标签 a 的用户数；$\sum_{i=1}^{n}\sum_{a=1}^{s} f_{ia}$ 表示标签总数，$\sum_{i=1}^{n} f_{ia}$ 表示标签 a 的总数。

由式(1)可以看出，用户选择的标签被用户选得少并且此标签占整个标签集合的比重越小，这样就能在一定程度上明确用户偏好，从而提高推荐效率。

2.2 融合时间系数的衰减函数

传统的推荐算法对用户标签偏好常用静态标签标识，一般用 0 和 1 来表示。这样可以明显看出在任何时候这些标签所起到的推荐作用都是相同的，对于某些时效性较强的推荐并不能起到较好的推荐效果。例如：某用户以前喜欢古典音乐，现在喜欢流行音乐，如果不考虑用户兴趣偏好随时间变化就会导致推荐不贴合用户偏好[14]。在实际中用户的兴趣往往是处于动态变化中的[15]。相对于早期的用户行为，近期的用户行为对于推荐更有意义，因此将用户近期的标签给予较高的权重，从而使推荐更具有时效性，提高推荐效率。本文引入一种衰减函数并且融入时间系数来充分贴合用户兴趣偏好随时间的变化，公式如下：

$$T_{ui}=\mathrm{e}^{-\frac{lbT_s\times\left|\frac{t_{\mathrm{now}}-t_{ui}}{T_s}\right|}{T_{\mathrm{att}}}} \tag{2}$$

其中，$T_{ui}\in(0,1)$，代表用户 u 对项目 i 的时间权重；T_s 表示时间窗口参数，其值表示用户偏好兴趣持续时间；t_{now} 表示当前做推荐的时间，t_{ui} 表示用户对项目作出评价的时间；T_{att} 是时间衰减参数，代表兴趣偏好衰减速率；$\left|\frac{t_{\mathrm{now}}-t_{ui}}{T_s}\right|$ 表示对计算结果进行上舍入处理，$T_s\times\left|\frac{t_{\mathrm{now}}-t_{ui}}{T_s}\right|$ 表示用户评价项目时间所处的时间段。若用户在一周的时间内兴趣偏好基本没变，则认为该用户兴趣保持稳定的周期为 7 天，即 $T_s=7$。若用户评价完项目后在 7 天内进行推荐，即 $t_{\mathrm{now}}-t_{ui}\leqslant 7$，则用户兴趣在第 8 天后才开始衰减，每 7 天为一个衰减周期，衰减周期内衰减系数相同。

2.3 计算用户偏好相似性

根据前文分析，在利用 TF-IDF 方法计算用户兴趣偏好时加入融入时间系数的衰减函数得出用户兴趣偏好，更新用户标签矩阵中的值，公式如下：

$$P_{ua}=\frac{\sum_{i=1}^{n} r'_{ui}\times f_{ia}\times T_{ui}}{\sum_{i=1}^{n}\sum_{a=1}^{s} r'_{ui}\times f_{ia}}\times\lg\left(\frac{\mathrm{num}_m}{\mathrm{num}_{ua}}\times\frac{\sum_{i=1}^{n}\sum_{a=1}^{s} f_{ia}}{\sum_{i=1}^{n} f_{ia}}\right) \tag{3}$$

最后归一化欧式距离，公式如下：

$$\mathrm{sim}_1(u,v)=\frac{1}{1+\sqrt{\sum_{i=1}^{n}(u_i-v_i)^2}} \tag{4}$$

2.4 融合用户属性相似度

在计算相似度时，采用常规的相似的算法不会将不同用户的个人属性进行相似性对比，如性别和年龄等属性。因此，本文考虑了上述用户属性，并且将这些基本的用户属性融入到相似度计算中。

(1)年龄属性相似度，公式如下：

$$N(u,v)=\mathrm{e}^{-|n_u-n_v|} \tag{5}$$

其中，u 和 v 分别代表两个用户，$N(u,v)$ 的取值范围为 $[0,1]$ 之间，值越小相似度越小；n_u 和 n_v 分别为用户 u 和 v 的年龄。

(2)性别属性相似度，公式如下：

$$X(u,v)=\begin{cases}0, X_u\neq X_v\\ 1, X_u=X_v\end{cases} \tag{6}$$

其中，u 和 v 代表不同的用户，X_u 和 X_v 分别是用户 u 和 v 的性别。

(3)根据上述用户性别和年龄属性相似度，根据实际情况分别给予不同的权重得出用户属性相似度，公式如下：

$$\mathrm{sim}_2(u,v)=\alpha N(u,v)+(1-\alpha)X(u,v) \tag{7}$$

其中，权重系数 $\alpha\in[0,1]$，在不同的推荐场景和领域中可以根据实际情况对 α 值进行调整。

3 融合用户兴趣的协同过滤推荐算法

首先通过对 $\mathrm{sim}_1(u,v)$ 和 $\mathrm{sim}_2(u,v)$ 线性组合，将用户兴趣偏好和属性融合得到综合相似度，得到一种新的相似度计算模型，公式如下：

$$\mathrm{sim}(u,v)=\lambda\mathrm{sim}_1(u,v)+(1-\lambda)\mathrm{sim}_2(u,v) \tag{8}$$

式中，$\lambda\in[0,1]$ 为权重系数，$\mathrm{sim}(u,v)$ 值与两个用户的相似性成反比关系。

然后对项目进行评分预测，最后进行推荐，公式如下：

$$P_{ui}=\overline{r_u}+\frac{\sum_{v\in N_u}\mathrm{sim}(u,v)\times(r_{vi}-\overline{r_v})}{\sum_{v\in N_u}|\mathrm{sim}(u,v)|} \tag{9}$$

其中，$\overline{r_u}$表示用户 u 给评价项目的平均分，$\overline{r_v}$表示近邻用户 v 评价项目的平均分，N_u 表示目标用户 u 的最近邻居，v 表示邻居集合中对项目有评分的用户，r_{vi} 表示用户 v 对项目 i 的评分，$\mathrm{sim}(u,v)$表示用户 u、v 的相似度。

4 实验与分析

4.1 实验数据

实验采用开源的数据集 MovieLens-1M。实验中使用交叉验证方式对用户评分进行预测。

4.2 评估指标

经过多轮训练减小评分误差，获得最优参数推荐模型。常用评价指标是平均绝对误差(MAE)，这种误差计算方式见式(10)：

$$\mathrm{MAE}=\frac{\sum_{u,i\in \mathrm{Test}}|P_{ui}-r_{ui}|}{\sum_{u\in \mathrm{Test}}|\mathrm{Test}|} \tag{10}$$

其中，r_{ui} 为用户 u 对项目 i 的真实评分，P_{ui} 为用户 u 对于项目 i 的预测评分。分母为测试集，分子为用户 u 对项目 i 真实评分和预测分数的差值。通过计算 Test 中 P_{ui} 与 r_{ui} 的平均绝对误差，评估模型的性能。

4.3 结果分析

首先确定本文涉及到的参数值，参数分别为：T_s、T_{att} 和 λ。

实验 1：通过 MAE 值来确定时间窗口参数 T_s 的值。如图 2 所示，在 $K=50$ 时，$T_{att}=20$、$T_{att}=40$、$T_{att}=60$、$T_{att}=80$、$T_{att}=100$ 的条件下，MAE 的值的变化趋势都是先降后升。当 $T_{att}=40$，$T_s=4$ 时，MAE 值最小；当 $T_{att}=100$，$T_s=5$ 时，MAE 值最小；当 T_{att} 分别为 20、60 和 80，$T_s=6$ 时，MAE 值最小。令 $T_s=6$ 来进行后续的实验，即用户的兴趣偏好的变化周期为 6 天。

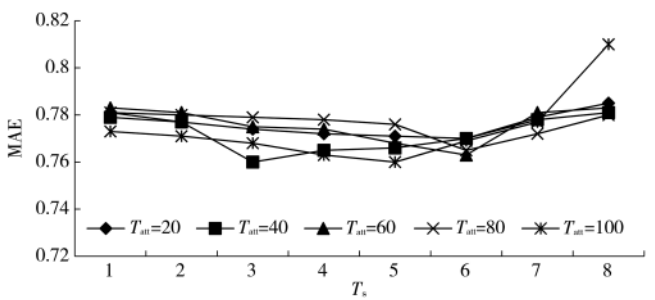

图 2 不同 T_s 值对应的 MAE 值

实验 2：判定 T_{att} 的值。如图 3 所示，在 $K=50$，$T_s=6$ 时，$T_{att}=30$、$T_{att}=40$、$T_{att}=50$、$T_{att}=60$、$T_{att}=70$、$T_{att}=80$、$T_{att}=90$ 时，MAE 的值先下降；到 $T_{att}=60$ 时，MAE 值达到最低，然后上升。所以令 $T_{att}=60$，进行后续实验。

实验 3：确定式(8)中参数 λ 的值。当 $\lambda=1$ 时，$\mathrm{sim}(u,v)=\mathrm{sim}_1(u,v)$，表示只利用用户的兴趣偏好来计算用户之间的相似性；当 $\lambda=0$ 时，$\mathrm{sim}(u,v)=\mathrm{sim}_2(u,v)$，表示仅利用

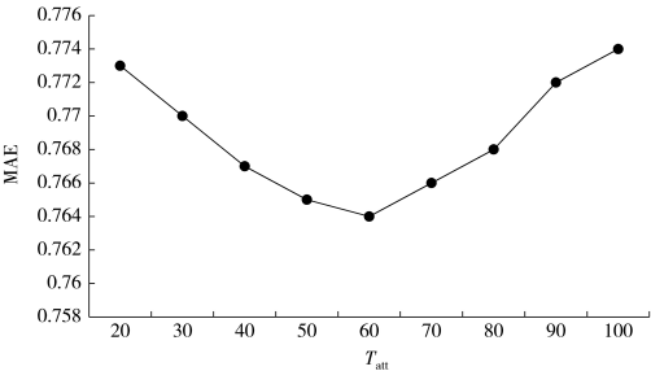

图 3 不同 T_{att} 值对应的 MAE

用户的属性计算用户之间的相似性。如图 4 所示，在 $K=20$、$K=40$、$K=60$、$K=80$ 时，MAE 值先下降后上升；当 $\lambda=0.4$ 时，MAE 值最小，推荐效果最好。

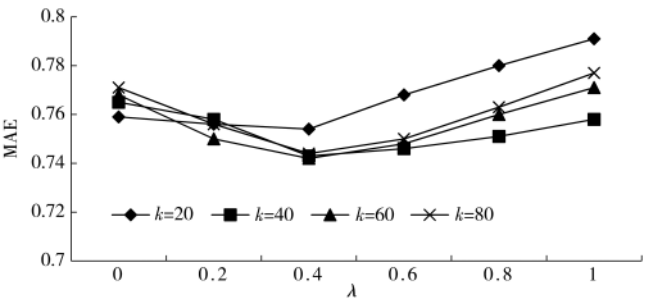

图 4 不同 λ 对应的 MAE 值

实验 4：在近邻不同的情况下，比较了不同推荐算法的推荐性能，其中包括将基于用户的协同过滤推荐算法(UBCF)[16]、基于 K-means 聚类的协同过滤推荐算法(K-means UBCF)[17]、基于 Canopy+K-means 混合聚类的协同过滤推荐算法(Canopy+K-means UBCF)与本文提出的算法进行了对比。得出的实验结果如图 5 所示。

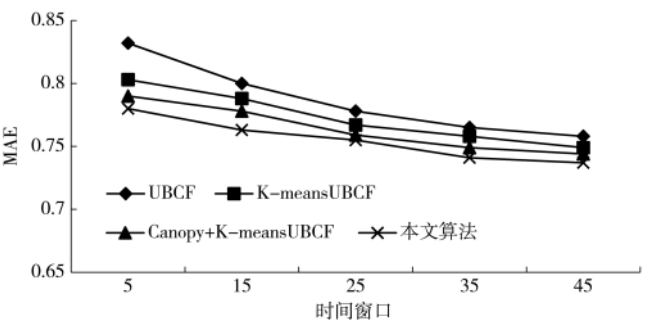

图 5 不同算法对应的 MAE 值

由图 5 可知，随着目标用户最近邻居个数的增加，实验中所用的 UBCF、K-means UBCF、Canopy+K-means UBCF 和本文所提出的算法的 MAE 值都会逐渐降低并趋于一个稳定值。由图 5 可以直观地发现，本文所提出的算法相对于其他 3 种算法推荐准确度最高。例如，

当最近邻居个数为35时，Canopy+K-means UBCF的MAE值为0.758，同样条件下本文所提出的算法的MAE值为0.741，推荐效果提升了1.7%。

5 结论

本文提出一种基于混合聚类与融合用户偏好的协同过滤推荐算法，通过建立Canopy+bi-Kmeans混合聚类模型并且将传统的相似性度量算法中加入用户属性和用户兴趣偏好。实验结果表明，本文提出的基于混合聚类与融合用户偏好的协同过滤推荐算法在一定程度上提高了推荐可靠性。由于本文的算法是在各方面条件较为理想的环境下实现的，其鲁棒性和稳定性有待提高，因此下一步的工作是将该算法运用到现实项目中，并且不断追求更高的推荐效率。

参考文献

[1] 陆航，师智斌，刘忠宝.融合用户兴趣和评分差异的协同过滤推荐算法[J].计算机工程与应用，2020，56(7)：24-29.

[2] LINDEN G，SMITH B，YORK J.Amazon.com recommendations：item-to-item collaborative filtering[J].Internet Computing.IEEE，2003，7(1)：76-80.

[3] CELMA Ò，SERRA X.Foafing the music：bridging the semantic gap in music recommendation[J].Web Semantics Science Services & Agents on the World Wide Web，2008，6(4)：250-256.

[4] HOPFGARTNER F，BRODT T，SEILER J，et al.Benchmarking news recommendations[J].ACM SIGIR Forum，2016，49(2)：129-136.

[5] 常江.基于Apache Mahout的推荐算法的研究与实现[D].成都：电子科技大学，2013.

[6] 王蓉，刘宇红，张荣芬.基于混合聚类与融合用户属性特征的协同过滤推荐算法[J].现代电子技术，2021，44(6)：179-182.

[7] 赵伟，林楠，韩英，等.一种改进的K-means聚类的协同过滤算法[J].安徽大学学报(自然科学版)，2016，40(2)：32-36.

[8] 张琳，牟向伟.基于Canopy+K-means的中文文本聚类算法[J].图书馆论坛，2018，38(6)：113-119.

[9] 王林，贾钧琛.基于改进Canopy-K-means算法的并行化研究[J].计算机测量与控制，2021，29(2)：176-179，186.

[10] 魏佳代.基于DNS日志的用户访问行为分析和研究[D].北京：北京交通大学，2019.

[11] 郝雅娴.K-Means聚类中心最近邻推荐算法[J].山西师范大学学报(自然科学版)，2021，35(1)：72-78.

[12] LARSEN B，AONE C.Fast and effective text mining using linear-time document clustering[C]//Proceedings of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.ACM，1999：16-22.

[13] D′ADDIO R M，DOMINGUES M A，MANZATO M G.Exploiting feature extraction techniques on users′ reviews for movies recommendation[J].Journal of the Brazilian Computer Society，2017，23(1)：1-7.

[14] YAN H，PENG Q，HU X，et al.Web service recommendation based on time series forecasting and collaborative filtering[C]//2015 IEEE International Conference on Web Services.IEEE，2015：233-240.

[15] 涂金龙，涂风华.一种综合标签和时间因素的个性化推荐方法[J].计算机应用研究，2013，30(4)：1044-1047.

[16] 王卫红，曾英杰.基于聚类和用户偏好的协同过滤推荐算法[J].计算机工程与应用，2020，56(3)：68-73.

[17] 肖文强，姚世军，吴善明.基于用户谱聚类的Top-N协同过滤推荐算法[J].计算机工程与应用，2018，54(7)：138-143.

(收稿日期：2021-08-22)

作者简介：

麻天(1995-)，男，硕士研究生，主要研究方向：知识图谱与推荐算法。

余本国(1976-)，男，博士，副教授，主要研究方向：数据分析与深度学习。

扫码下载电子文档

(上接第28页)

the integration of time-sensitive communications in legacy LAN/WLAN[C]//2018 IEEE Globecom Workshops.IEEE，2018：1-7.

[10] DELPHI H，LUKMAN R，FITRI S R.Design and implementation of multi-protocol gateway for Internet of Things[C]//2019 IEEE International Conference on Industry 4.0，Artificial Intelligence，and Communications Technology(IAICT)，2019：64-69.

(收稿日期：2022-03-08)

作者简介：

王博(1996-)，男，硕士研究生，主要研究方向：工业以太网、集成电路设计。

王雪迪(1996-)，男，硕士研究生，主要研究方向：工业以太网、集成电路设计。

时广轶(1979-)，男，博士，教授，主要研究方向：MEMS惯性器件、碳纳米管传感器设计与应用技术。

严伟(1966-)，通信作者，男，博士，教授，主要研究方向：视频编解码、无线通信电路设计，E-mail：yanwei@ss.pku.edu.cn。

扫码下载电子文档

版权声明