文章编号：2095-2163(2021)06-0210-04 中图分类号：TP301.6 文献标志码：A

基于多特征的数字图书推荐算法

李 冬
（商丘职业技术学院，河南 商丘 476001）

摘 要：数字图书资源随着现代信息技术的高速发展日益丰富多样，从海量的数字图书中为读者提供高质量的推荐服务愈发重要。基于单一维度的推荐算法，在推荐的有效性上略显不足。因此，本文提出一种融合数字图书多项特征的推荐算法，通过权重实验对不同特征赋予不同的权重，并以此为基础建立推荐模型。实验结果表明，融合多特征数字图书推荐方法能有效提高推荐性能，获得了较好的推荐效果。

关键词：多特征；推荐算法；数字图书

Digital book recommendation algorithm based on multi-features

LI Dong
（Shangqiu polytechnic, Shangqiu Henan 476001, China）

【Abstract】 With the rapid development of information technology, digital book resources have become rich and diverse, high quality digital book recommendation service become more and more important for reader. It is not good enough on the effectiveness based on single dimension recommendation algorithm. So proposed a recommendation algorithm that combined multi-features, through weight experiment, give different weight for different feature ,and then constructed recommendation model. Experimental results show that combined multi-features digital Book recommendation algorithm can improve the recommendation performance, and get better efficiently.

【Key words】 multi-features; recommendation algorithm; digital book

0 引 言

伴随着信息技术的高速发展，数字媒体技术日新月异，大量的数字资源的诞生和普及，对数字资源服务也提出越来越高的要求，如何从海量的数字图书中，根据相关的数据信息，为读者提供高质量、差异化、个性化的图书推荐愈发重要。提高图书推荐的效率和准确率，提高读者的满意度、粘合度是各种数字图书平台努力的目标和方向。

基于各种算法建立起来的数字图书推荐系统是根据读者的个人偏好，提供差异化图书推荐的有效方法。算法是推荐系统高效、准确运行的基础和关键，目前推荐系统常用的算法有基于内容的推荐算法、基于知识的推荐算法、基于关联规则的推荐算法、基于协同过滤推荐算法以及基于模型的各类推荐算法[1]。以这些算法建立起来的推荐系统通过对用户历史行为数据的分析，得出用户的真实需求，向用户推荐相关的产品及信息，随着正反馈结果的不断提高，加强了用户和平台间的紧密度，实现用户链式反应增值，这些推荐系统在电子商务、音视频推荐、新闻、图书等很多领域已经取得的广泛的应用，产生了很好的经济效益和社会效益。

数字图书和普通图书相比在数据信息和数据质量上更加的丰富和准确，读者对数字图书的评价可以更加的便捷、有效，数字图书的名称、简介、评论、作者、出版社、出版时间、上线时间、搜索量、浏览频次、页面停留时间等因素都可能会影响读者的兴趣偏好。基于某一特征建立起来的推荐系统，在一定程度上欠缺了对其它影响因素的考虑，在推荐的有效性上略显不足。因此，本文提出一种融合数字图书多项特征的推荐算法，并以此为基础建立推荐模型。

1 多特征数字图书的数据处理

通过对多个数字图书管理系统中的数据研究发现，数字图书的数据属性主要有名称、简介、评论、作者、出版社、出版时间、读者信息、图书评分等等。找到合适的方法，融合这些数据，以此为基础构建数字图书的推荐方法，下面介绍各种数据特征的处理和模型构建。

基金项目：2018 年度河南省高等学校青年骨干教师培养计划项目(2018GGJS221)。

作者简介：李 冬(1982-)，男，硕士，副教授，主要研究方向：计算机应用技术。

收稿日期：2021-04-23

1.1　数字图书简介信息的数据处理及特征提取

数字图书简介信息主要采用文本展示，基于卷积神经网络 CNN 在文字识别中表现出较好的识别效果，并且对于未知样本的类标号也具有较好的预测性，本文采用卷积矩阵分解 ConvMF 的算法，对数字图书简介信息进行处理，得到数字图书预测评分矩阵 P_1。

忽略标点符号、空格等无效信息，通过 Word2Vec 模型计算得到数字图书简介信息的词向量矩阵，输入 CNN 中。每条数字图书的最大简介信息单词数为 max_*lenth* = 300，超出单词直接截断。所有数字图书简介信息单词形成序列 L，基于数据库中数据大小的考虑，选取出现最多的前2 000个单词组成列表 V_s，用 UNK 对应的词向量表示仅在 L 中出现的单词。数字图书简介信息组成 $m \times n$ 矩阵，m 为简介信息的单词序列，$\boldsymbol{n}$ 为每个单词向量维度；如若卷积神经网络输出的数字图书分类类别为未知，则视未知类别数字图书特征向量为 $\boldsymbol{V}_1$。

定义读者数量为 M，数字图书数量为 N，$\boldsymbol{U}_i$ 表示读者特征向量，$\boldsymbol{V}_j$ 表示数字图书特征向量，R_{ij} 表示读者 i 对数字图书 j 的评分，$\boldsymbol{W}$ 为卷积神经网络中的权重向量，W_k 为第 k 列元素，ε 表示读者整体评分矩阵 R 与读者、数字图书的特征向量内积之差的方差，ε_u、ε_v、ε_w 分别为读者特征向量矩阵 $\boldsymbol{U}$、数字图书特征矩阵 $\boldsymbol{V}$ 和卷积神经网络中内部权重 $\boldsymbol{W}$ 的方差。结合公式(1)，利用随即梯度下降法求解 $\boldsymbol{U}$ 和 $\boldsymbol{V}$。

$$E = \frac{1}{2}\sum_{i=1}^{M}\sum_{i=1}^{M} I_{ij}\,(R_{ij} - U_i^{\mathrm{T}}V_j)^2 + \frac{\xi u}{2}\sum_{i=1}^{M}|U_i|^2 + \frac{\xi_u}{2}\sum_{j=1}^{N}|V_j - V_1|^2 + \frac{\xi_w}{2}\sum_{k=1}^{W}|W_k|^2. \tag{1}$$

其中，$\xi_u = \frac{\varepsilon^2}{\varepsilon_u^2}$；$\xi_v = \frac{\varepsilon^2}{\varepsilon_v^2}$；$\xi_w = \frac{\varepsilon^2}{\varepsilon_w^2}$；$I_{ij} = 1$ 表示读者 i 对数字图书 j 作出过评价；$I_{ij} = 0$ 表示读者 i 对数字图书 j 未作出过评价。

卷积矩阵分解算法中引入概率模型优化矩阵分解，利用已知数据预测评分矩阵中的未知值，将上文得到数字图书特征向量 $\boldsymbol{V}_1$ 与矩阵概率分解相结合，能很好地预测读者对数字图书的预测评分 P_1，P_1 的取值在[0,5] 之间。

1.2　数字图书评论信息处理

读者对数字图书的评论会用许多带有感情色彩的词汇，这些词汇也是读者对图书喜爱程度的表达，对图书推荐具有重要的参考价值。因此，对这些图书评论中词汇进行量化处理，得到读者对数字图书的预测评分矩阵 $\boldsymbol{P}_2$。

用 AFINN 情感词典对图书评论中的情感词汇进行量化，每一个关键性词汇对应一个情感分值，取值范围在[-5,5]之间，经过处理计算可以得到每条评论的总情感分值[2]。利用 Python 自然语言工具包对评论语言进行分词，并根据 Natural Language Toolkit 中的停用词表，进行停用词过滤，建立结构化的评论数据[3]。

AFINN 情感取值介于[-5,5]之间，因此可以将正向积极的评论取值为(0,5]，负向消极的评论取值为(0,-5]，中性评价取值为 0，利用公式(2)计算得出总的情感分值。

$$G(Q_{ui}) = \frac{\sum_{wj \in Q_{ui} \& wj \in K} W(w_j)}{|w_j \in Q_{ui} \& w_j \in K|}. \tag{2}$$

其中，$Q_{ui} = (w_1, w_2, \cdots, w_j)$，$Q_{ui}$ 表示读者 u 对数字图书 i 的结构化评论；w_j 是第 j 个单词或词汇；$W(w_j)$ 是每个单词或词汇的情感分值；K 为 AFINN 中的词汇。

利用公式(3)对 $G(Q_{ui})$ 所得结果进行泛化处理，使其结果取值在[0,5]之间，$x \in [-5,5]$，$y \in [0,5]$，得到读者评论的图书预测评分矩阵 $\boldsymbol{P}_2$。

$$y = 0.5x + 2.5. \tag{3}$$

1.3　对图书作者和出版社进行数据建模

作者、出版社对于数字图书的评分也有着较高的影响力，因此将其作为影响图书最终预测评分的影响因子，赋予一定的权重。

最近邻方法 KNN 可以对一个不知类别的样本找出最相似的近邻用户进行分类，采用此方法求出近邻读者对作者 d_s 所有数字图书的评分均值 $\overline{\rho_{d_s}}$，以及近邻读者对出版社 e_o 所有数字图书的评分均值 $\overline{\lambda_{e_o}}$，利用公式(4) 计算出其均值，作为作者、出版社共同影响下，读者对数字图书 i 综合评分为 $P(i)\,d_s e_o$，表示作者为 d_s，出版社为 e_o，读者对图书 i 的综合评分。

$$P(i)\,d_s e_o = \frac{\overline{\rho_{d_s}} + \overline{\lambda_{eo}}}{2}. \tag{4}$$

根据 $P(i)\,d_s e_o$ 得出的结果，利用公式(5)可以构建读者 u 对图书 i 的评分预测矩阵 $\boldsymbol{P}_3$，$P'(ui)$ 为读者 u 对图书 i 的评分。

$$\boldsymbol{P}_3 = P'(ui) = \begin{bmatrix} p(1)\ d_1 e_1 & \cdots & p(h)\ d_1 e_o \\ \vdots & \ddots & \vdots \\ p(v)\ d_s e_1 & \cdots & p(i)\ d_s e_o \end{bmatrix}. \quad (5)$$

1.4 读者-图书评分数据的处理

基于读者、图书、图书评分矩阵,通过协同过滤技术进行图书推荐已相对成熟,无需对数据再进行特别的处理。根据数据源 $D=(U,I,R)$,结合协同过滤算法,利用余弦相似度计算,可以得到目标读者对图书的预测评分矩阵 $\boldsymbol{P}_4$,其中 $U=\{User_1,User_2,\cdots,User_i\}$ 为读者样本集合,$I=\{Item_1,Item_2,\cdots,Item_j\}$ 为数字图书样本集合,$\boldsymbol{R}$ 为 $i\times j$ 阶矩阵,是已有读者对各数字图书的实际评分矩阵。

2 融合多特征数字图书数据的模型构建

根据多特征数字图书的数据处理,重点研究了图书简介信息、图书评论、图书作者和出版社以及图书的评分等影响因子,以此为基础分别构建了读者对数字图书的预测评分矩阵 $\boldsymbol{P}_1$、$\boldsymbol{P}_2$、$\boldsymbol{P}_3$、$\boldsymbol{P}_4$,将每个影响因子赋予一定的权重,利用公式(6) 融合计算,作为最终预测评分 P_{ui}。

$$P_{ui} = \alpha\boldsymbol{P}_1 + \beta\boldsymbol{P}_2 + \gamma\boldsymbol{P}_3 + \delta\boldsymbol{P}_4. \quad (6)$$

其中,α、β、γ、δ 为不同预测评分矩阵相应的权重,并且 $\alpha+\beta+\gamma+\delta=1$,通过问卷调查的方式获取图书简介信息、图书评论、图书作者和出版社以及图书的评分等因素对读者选择图书的直观影响程度,根据问卷结果,设定 α、β、γ、δ 的初始值,不断调整权重,对不同的权重组合进行比较,取最小的 *MAE* 值所对应的 α、β、γ、δ 值作为公式中的权重值。

P_{ui} 为读者 u 对图书 i 综合多特征的预测评分,根据前文所述,P_1 为读者 u 根据图书简介信息对图书 i 的预测评分;P_2 为读者 u 根据图书评论对图书 i 的预测评分;P_3 为读者 u 根据图书作者和出版社对图书 i 的预测评分;P_4 为读者 u 根据图书的评分对图书 i 的预测评分,$P_1,P_2,P_3,P_4\in[0,5]$。根据已经确定的 α、β、γ、δ 权重值分别赋予 P_1、P_2、P_3、P_4,$\alpha P_1+\beta P_2+\gamma P_3+\delta P_4$ 所得结果即为 P_{ui},得到目标读者对 未选择图书的综合预测评分后,根据评分由高到底排序,将评分最高的前 k 个图书推荐给该读者。

3 实验分析

3.1 实验数据

本文采用商丘市图书馆数字资源管理中心提供的数据集开展实验,数据集记录了 2020 年 9 月 1 日~2021 年 1 月 10 日期间 681 位读者对2 936本图书的152 600 个评分记录,提供的数据信息包括了读者 ID、性别、年龄、专业以及数字图书 ID、名称、作者、出版社、图书简介、读者评论、评分等。根据实验需要,利用爬虫程序获取数字图书简介、读者评论、作者、出版社、读者评分等数据进行训练和实验检验。

3.2 评价指标

平均绝对偏差 *MAE*(Mean Absolute Error)体现预测评分与真实评分之间的偏差平均值,计算公式如式(7)所示:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|p_i - r_i|. \quad (7)$$

其中,n 为读者数量;P_i 为预测读者评分集合 $\{p_1,p_2,\cdots,p_N\}$;r_i 为实际读者评分集合$\{r_1,r_2,\cdots,r_N\}$;计算出的 *MAE* 值越小,误差越小,推荐效果越好。

3.3 实验结果及分析

首先进行权重调整实验,获得最佳的权重组合对数字图书的评分矩阵 $\boldsymbol{P}_1$、$\boldsymbol{P}_2$、$\boldsymbol{P}_3$、$\boldsymbol{P}_4$ 权重赋值,然后验证融合多特征数字图书推荐性能。

3.3.1 权重调整实验

权重 α、β、γ、δ 取值组合范围较大,通过对 50 位读者直观感受和实际经验进行的问卷调查显示,数字图书简介信息、读者评论、评分对其选择图书的影响较大,因此可以假定数字图书简介信息、读者评论、评分对图书推荐结果的影响较大,作者、出版社对图书推荐结果的影响较小,设置初始值 $\alpha=0.3$,$\beta=0.3$、$\gamma=0.3$、$\delta=0.1$,不断调整权重进行测试。邻居数 N 在 10 - 50 之间取值,当 N 取值 30 时,不同权重对应的 *MAE* 值见表 1。

表 1 $N=30$ 时不同权重对应 *MAE* 值

Tab. 1 $N=30$, Different weights correspond to *MAE* values

编号	权重设置				*MAE*
	α	β	γ	δ	
1	0.3	0.3	0.1	0.3	0.84
2	0.3	0.2	0.2	0.3	0.85
3	0.2	0.3	0.2	0.3	0.82
4	0.2	0.3	0.1	0.4	0.79
5	0.4	0.2	0.1	0.3	0.81
6	0.4	0.3	0.1	0.2	0.83

实验结果如图 1 所示,权重编号为 4、12、19 时 *MAE* 值较小,采用权重编号 4 所对应的权重,取值 $\alpha=0.2$、$\beta=0.3$、$\gamma=0.1$、$\delta=0.4$ 进行后续的数字图书推荐实验。

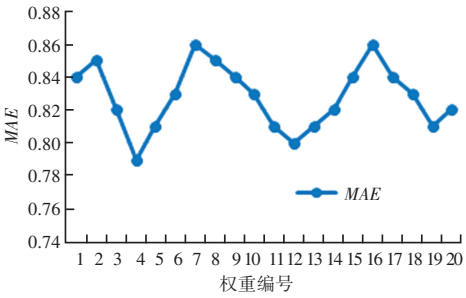

图 1　$N=30$ 时不同权重编号对应的 *MAE* 值

Fig. 1　$N=30$, Different weights serial number correspond to *MAE* values

3.3.2　融合多特征数字图书推荐性能实验

该实验验证本文提出的融合多特征数字图书推荐性能，用协同过滤算法 CF 与本文提出的方法进行对比，比较平均绝对偏差 *MAE* 值。协同过滤算法 CF 得到的预测评分矩阵就是目标读者对图书的预测评分矩阵 $\boldsymbol{P}_4$，得出的 *MAE* 值如图 2 所示。

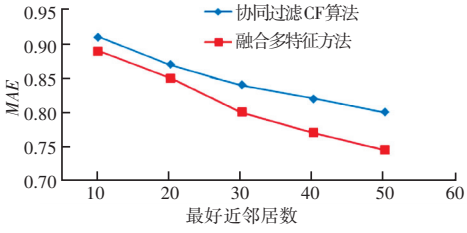

图 2　最近邻居数变化时对应的 *MAE* 值

Fig. 2　*MAE* values of nearest neighbors' number changes

实验表明，融合多特征数字图书推荐方法与协同过滤 CF 算法相比较，*MAE* 值均最小，表明本文提出的数字图书推荐方法的有效性，该方法在一定程度上提高了数字图书的推荐性能，获得了较好的推荐效果。

4　结束语

数字图书具有多特征属性，随着现代信息技术的发展，数字图书特征数据已经极大的丰富，这为融合多特征数字图书推荐奠定了基础。本文通过对数字图书特征的分析，考虑图书简介、读者评论、作者、出版社、读者评分等多种影响因素，分别对图书评分进行预测，对预测结果加权融合，赋予一定的权重，以此提高图书的推荐性能。通过实验证明该方法优于协同过滤 CF 算法，具有更好的数字图书推荐性能。

参考文献

[1] PAZZANI M J, BILLSUS D. Content-based recommendation systems[M]. The Adaptive Web. Springer Berlin Heidelberg, 2007:325-341.

[2] HANSEN L K, ARVIDSSON A, NIELSEN F A, et al. Good friends, bad news - Affect and virality in Twitter [J]. Communications in Computer & Information Science, 2011, 185: 34-43.

[3] 李晨瑞. 基于语义推理和表示的机器阅读理解研究[D]. 上海：华东师范大学，2018.