李宏志,李苋兰. 融合改进的内容与协同过滤的博客推荐方法[J]. 湖南科技大学学报(自然科学版), 2021, 36(3):104-112. doi:10.13582/j.cnki.1672-9102.2021.03.015

LI H Z, LI X L. Recommendation Method by Fusion of Improved Content-Based Recommendation and Collaborative Filtering [J]. Journal of Hunan University of Science and Technology (Natural Science Edition), 2021, 36(3):104-112. doi:10.13582/j.cnki.1672-9102.2021.03.015

融合改进的内容与协同过滤的博客推荐方法

李宏志[1*],李苋兰[2]

(1.上海海事大学 信息工程学院,上海 200000;
2. 安徽省计量科学研究院 新能源汽车产业中心,安徽 合肥 230000)

摘 要:在中文博客系统中,受限于用户特征信息的稀少,使用协同过滤算法的准确率往往不高,而基于内容推荐算法,又会影响推荐结果的多样性.因此,文章提出了一种融合改进的内容推荐与协同过滤相结合的推荐方法.首先,采用协同过滤算法发现用户的潜在兴趣并通过谱聚类改进协同过滤的相似度计算,提高处理效率;其次,基于改进的内容的推荐算法构建用户的既有兴趣模型,计算潜在推荐内容与既有兴趣模型的匹配度;最后,通过逻辑回归算法融合协同过滤与内容推荐的结果.实验结果显示,文章所提出的推荐方法相对于单一的协同过滤和内容推荐可以显著提高推荐的结果的准确率和召回率,具备良好的推荐效果.

关键词:谱聚类;协同过滤;逻辑回归;基于内容的推荐

中图分类号:TP311 **文献标志码**:A **文章编号**:1672-9102(2021)03-0104-09

Recommendation Method by Fusion of Improved Content-Based Recommendation and Collaborative Filtering

LI Hongzhi[1], LI Xianlan[2]

(1. College of Information Engineering, Shanghai Maritime University, Shanghai, 200000, China;
2. New Energy Vehicle Industry Center, Anhui Institute of Metrology, Hefei 230000, China)

Abstract: In the use of the blog system, it is difficult to use content-based recommendation algorithm subject to rare user feature information, and the diversity of recommendation results will be affected when only the content-based recommendation was used . A new method of blog recommendation based on fusion of improved content-based recommendation and collaborative filtering was proposed. Firstly, the collaborative filtering algorithm was used to discover the potential interest points of users, and the similarity calculation of collaborative filtering was improved through spectral clustering. Secondly, the existing interest model of users was constructed by means of content-based recommendation method. Finally, the comprehensive result of collaborative filtering and the improved content-based recommendation was determined by logistic regression. Compared with contented-based recommendation and collaborative filtering through experiment, the accuracy and recall rate of recommended results re both improved significantly.

Keywords: spectral clustering;collaborative filtering;logistic regression;content-based

收稿日期:2019-01-14
基金项目:安徽省自然科学基金资助项目(1408085MF126)
* **通信作者**,E-mail: 1071260932@qq.com

随着互联网的发展,博客系统成为当前重要网络的社交媒体之一,具有用户活跃度高、实时性强、信息量大等特点[1].但同时博客的信息过载的问题也日趋严重,令用户无从选择,因此,各类推荐系统被广泛地应用于博客系统中帮助用户快速准确地定位自己感兴趣的内容.目前,文凯和朱传亮等提出从用户的信任关系角度衡量网页或博客关系的重要性,将信息检索技术中的 PageRank 算法应用于博客内容的推荐[2];Hung C 等提出利用特征工程的相关算法应用于内容推荐研究,通过使用合适的数据挖掘与分析技术提炼出来的特征比原始特征具备更好的推荐效果[3];杨琛等提出了用户兴趣模型的概念,通过提取用户的历史记录构建特定的兴趣模型,利用兴趣模型计算出内容的匹配指数;但在内容推荐的多样性上存在不足,难以发掘用户的潜在兴趣[4-7];Panigrahi S 等提出了使用协同过滤算法进行内容推荐,通过用户的行为记录得出用户们的相似度矩阵,利用用户们的行为相似性向目标用户推荐内容[8-11],但协同过滤存在着系统冷启动问题,推荐的准确度依赖于用户行为记录的详细程度,通常需要经过一定量的用户记录才能形成推荐.随着社交媒体的发展,可以借助微博、Twitter 等社交媒体,加强内容算法在热点事件上的反应能力[12-13];林杰等提出了基于信任关系与语义分析的博客推荐算法[14],利用社交网络上信任关系提高推荐结果的准确性.杨武等提出了融合推荐模型,总体上采用协同过滤算法以内容推荐作为信息补充[15].总结起来,目前常用的博客推荐方法主要有基于文本内容的推荐方法、基于关联规则的推荐方法以及基于协同过滤算法的混合推荐方法[16].协同过滤算法是推荐系统中被应用的最早也最为广泛的算法之一,但基于协同过滤的算法存在系统冷启动的问题并且需要面对较为严重的数据稀疏性问题[17];而基于内容的推荐算法容易造成兴趣单一,马太效应明显.

针对上述问题,本文提出一种基于内容和协同过滤相融合的推荐方法,该方法改进了传统的基于内容推荐方法得到用户兴趣模型,提出了一种模型更新算法,解决推荐的局部性问题.对于传统的协同过滤方法,使用谱聚类算法挖掘更多的用户关联性;减少了相似度计算的复杂度;最后提出了基于逻辑回归的融合方法,将 2 种推荐结果有效的融合起来,用于最终的推荐得分.相对于单纯地依靠基于内容或协同过滤进行推荐,本文所提出的算法在保持推荐的召回率和准确率方面都有明显的优势.

1　改进的内容推荐算法

基于内容推荐算法,其基本思想是利用待推荐用户的以往行为记录,建立基于用户自身兴趣的推荐模型,为用户推荐与兴趣模型匹配度最高的内容.该算法的优点在于简单、高效,直接为用户推荐自己喜欢的内容;但缺点也很明显,推荐内容的广度不够,用户容易局限于自己的“小圈子”内,视野狭窄,同时兴趣模型的准确与否直接决定了推荐的效果,因此本文提出了一种兴趣淘汰算法,用于更新兴趣模型.

1.1　文本内容向量化

推荐系统中待推荐内容需要经过预处理才能被计算机处理,这一过程也被称作内容标签化.对于主要由文字内容呈现的博客的系统,其内容的标签化过程:

定义 1　主要特征词序列.给定博客集 $\boldsymbol{D}=\{d_1,d_2,\cdots,d_n\}$,将能够描述博客文本内容的关键词称为博客集的主要特征词序列 $\boldsymbol{T}=\{t_1,t_2,\cdots,t_k\}$,其中 k 代表了特征词的数目.

对于给定的博客内容 $\boldsymbol{D}=\{d_1,d_2,\cdots,d_n\}$ 和主要的特征词序列 $\boldsymbol{T}=\{t_1,t_2,\cdots,t_k\}$,博客内容 d_i 需要表示为计算机能够直接处理的向量空间(Vector Space Model,VSM),$d_i=\{w_{i1},w_{i2},\cdots,w_{ik}\}$,其中 w_{ij} 表示特征词 t_j 在文档 i 中出现的频率.文中权重采用 TF-IDF(term frequency-inverse document frequency)算法计算,如式(1)所示.

$$w(d,t_i)=[\mathrm{tf}\ (d,t_i)\ln(N/n+1)]/\sum_{t_i\in d}[\mathrm{tf}\ (d,t_i)\ln(N/n+1)]. \tag{1}$$

式中:$w(d,t_i)$ 为特征项 t_i 在文档 d 中的权重;$\mathrm{tf}\ (d,t_i)$ 为特征项 t_i 在文档中的词频;N 为训练文本的数量;n 为特征项 t_i 在样本中出现的次数.

1.2 兴趣模型的构建

用户兴趣模型的建立是内容推荐算法的关键步骤,基本思想是根据用户的历史活动记录,获取个人偏好;通常采用基于概率统计的方法,一定范围内的历史记录中出现频率越高的标签项,用户对于其兴趣程度也就越高,由此按照优先级建立由表征兴趣的关键词,组成用户的兴趣模型.兴趣模型的建立过程可划分为以下 3 个主要阶段:

1)用户历史数据的预处理阶段,如图 1 所示,这一阶段主要包括用户浏览记录的清洗,筛选出数据质量较高的记录,对有效的用户记录数据进行标准化和规格化处理,并经过主要特征词序列 $\boldsymbol{T}$ 的筛选,最终形成能够供计算机处理的向量空间模型 $\mathbf{DM}_u$ (Dataset-Model),对 $\mathbf{DM}_u$ 的定义为

$$\mathbf{DM}_u = \begin{bmatrix} \boldsymbol{\alpha}_{u1} \\ \boldsymbol{\alpha}_{u2} \\ \vdots \\ \boldsymbol{\alpha}_{uk} \end{bmatrix}. \tag{2}$$

式中:$\boldsymbol{\alpha}_{ui}$ 为用户 u 的历史记录 i 在经过处理后的主要特征词序列 $\boldsymbol{T}$ 的权值向量.

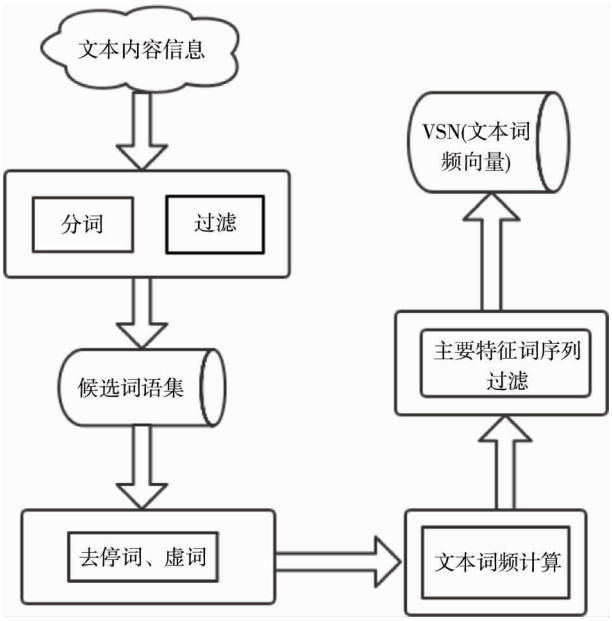

图 1 博客内容标签化过程

2)模型的建立阶段,在这一阶段中利用预处理阶段的获得的向量空间模型 $\mathbf{DM}_u$,对数据集中各主要特征词进行权重进行规约合并,形成用户的当前时刻兴趣模型 $\mathbf{IM}_u$ (Interest-Model),对于 $\mathbf{IM}_u$ 定义为

$$\mathbf{IM}_u = \{\beta_{u1}, \beta_{u2}, \cdots, \beta_{uk}\}. \tag{3}$$

式中:β_{ui} 为用户的历史数据文档在主要特征词序列 $\boldsymbol{T}$ 中的 t_i 处的文本权重,显然 β_{ui} 的计算需要考虑时间因素,因此引入时间影响因子.

定义 2 时间影响因子.设用户 u 的阅读的博客集为 $\boldsymbol{D} = \{d_1, d_2, \cdots, d_n\}$,其中 d_i 的阅读时刻为 h_i,当前时刻为 h_{now},则定义新闻 d_i 的时间影响因子如式(4)所示.

$$u_i = \frac{\ln(|h_i - h_{\text{now}}|^{-1})}{\sum\limits_{d_j \in \boldsymbol{D}} \ln(|h_j - h_{\text{now}}|^{-1})}. \tag{4}$$

对于 $\mathbf{IM}_u$ 中的 β_{ui} 的计算方法如式(5)所示.

$$\beta_{ui} = \sum_{t_i \in \boldsymbol{T}, d_j \in \boldsymbol{D}} (u_j w_{ij}). \tag{5}$$

式中:w_{ij} 为特征词 t_i 在文档 d_j 中的权重;u_j 为 d_j 的时间影响因子.

3)模型的更新阶段,在网络中用户的个人兴趣,会随着时间的迁移而不断变化的,因此用户的兴趣模型 IM_u 也是一个与时间相关的概念,需要模型能够根据用户行为的反馈不断地更新,以保证准确地反映用户的兴趣内容.

兴趣模型的构建过程如算法 1 所示.

算法 1 **UIM**(User-Interset-Model)构建算法

输入 用户阅读的博客内容集权值矩阵 $\mathbf{DM}_u$,主要特征词向量 $\boldsymbol{T}$,当前的时刻 h_{now},内容 d_i 的阅读时刻 h_i;

输出 用户的兴趣模型 $\mathbf{IM}_u$.

Begin

定义用户兴趣模型并初始化 $\mathbf{IM}_u \leftarrow \{0,0,\cdots,0\}$;

Step1　计算文档集中各文档时间影响因子;

foreach $i \in \text{rows}(\mathbf{DM}_u)$:

$$u_i \leftarrow \frac{\ln(|h_i - h_{\text{now}}|^{-1})}{\sum_{d_j \in D} \ln(|h_j - h_{\text{now}}|^{-1})}$$

endforeach;

Step2　根据 **Step1** 的时间影响因子,对 $\mathbf{DM}_u$ 中特征词向量加权求值;

foreach　$t_j \in \boldsymbol{T}$:

foreach　i ∈ rows ($\mathbf{DM}_u$):

β_j += $u_i * w[i][j]$

endforeach

endforeach

Step3　根据 **Step2** 的计算结果,更新模型向量 $\mathbf{IM}_u$,并输出.

foreach　$t_j \in \boldsymbol{T}$: $\mathbf{IM}_u \leftarrow \beta_j$

End

1.3　模型更新算法

用户兴趣模型的建立依赖于用户的历史浏览记录,但用户的个人兴趣会随着时间的变迁而变化,根据文献[13-15]描述可知用户的个人兴趣容易受到网络上的热点事件的影响;因此兴趣模型需要及时更新,适时淘汰过时的兴趣点,才能保证推荐的及时性和准确性,本文提出了一种兴趣淘汰算法,算法的主要思想是利用历史数据建立模型时考虑到时间影响因素且兼顾用户的推荐反馈信息,算法的主要流程如图 2 所示,具体内容:

1)首先需要设置有效历史记录的时间阈值,选定阈值内的记录作为模型建立主要依据,阈值可以采用系统当前时间与阅读时间的差值.

2)兴趣模型的更新触发机制,模型的更新依赖于用户对推荐结果的反馈,用户对推荐列表内容的点击率作为衡量推荐结果的标准,将推荐点击率作为触发模型更新的阈值.

3)兴趣模型的更新,需要考虑用户对推荐的反馈,在新建模型的过程中根据推荐的反馈结果,去除点击率较低内容及与其相似的内容记录.

兴趣模型更新算法如算法 2 所示.

算法 2　**REUIM** 算法

输入　推荐点击率阈值 γ,更新时间阈值 λ,低点击率文本相似度 η,当前的时刻 h_{now};

输出　更新后的兴趣模型 $\mathbf{IM}_u$.

Begin

Step1　读取当前模型的用户反馈数据集 $\mathbf{FD}_u$,计算平均推荐点击率 k;

Step2　if ($k > \gamma$):结束更新;else :执行 **Step3**;

Step3　读取用户浏览内容的历史记录 $\mathbf{RD}_u$,按照如下规则筛选记录:

(1)记录的阅读时间 h_i ,且满足;//保证记录的时效性

(2) $h_i - h_{\text{now}} < \lambda$ 选取 $\mathbf{FD}_u$ 中所有点击率低于 $k/3$ 的记录组成内容集合 $\mathbf{LFD}_u$;

(3)对于 $\mathbf{RD}_u$ 中的记录需满足与 $\mathbf{LFD}_u$ 任意记录的相似度小于 η,即 $\text{sim}(\text{rd}_i, \text{lfd}_i) < \eta$;

Step4　将文本记录集 $\mathbf{RD}_u$ 转换成空间向量模型,并执行算法 1 得到最新的兴趣模型 $\mathbf{IM}_u$;

End

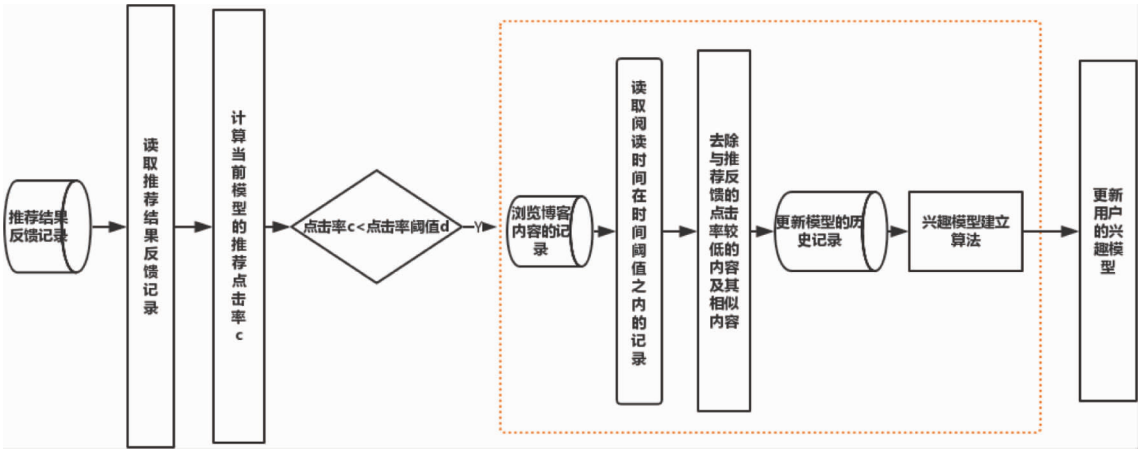

图 2 兴趣模型更新算法实现的流程图

2 基于谱聚类的协同过滤

2.1 谱聚类

通常,基于协同过滤的推荐算法需要计算用户或者物品之间的相似度[18];以基于用户的协同过滤为例,将相似度最高的 k 个元素组成相似用户集,以同类的用户的兴趣或者喜好为依据推荐内容,因此随着系统中用户数量的增加,推荐算法的执行效率会呈现下降的趋势;使用聚类算法用户根据相似度划分为不同的簇就显得非常有必要了,让用户相似度较高的聚集为同一簇,而不同用户簇之间保持较低的相关性[19].在进行推荐时,仅需要考虑所属簇的运算,减少了大规模的排序运算,缩小了查找范围.

谱聚类的算法思想源自于图论中关于图谱的划分的思想,其本质是将聚类问题转化为图的最优划分问题.在谱聚类的过程中,样本数据作为图中顶点 V,根据样本间的相似度将顶点间的边 E 赋值为权重 w,以此构成一个无向加权图 $G(V,E)$,数据的聚类问题就直接转化为图的切割问题.相比较 k-means 等传统聚类算法能够解决凸样本空间的局部最优问题,可在任意形状的样本空间进行聚类,具有更好的聚类效果[20].谱聚类的实现方法很多,但主要几个关键步骤可以归纳为

1) 针对用户数据集进行筛选、清洗等预处理工作,并计算相似度,构造相似度矩阵 $\boldsymbol{S} \in \boldsymbol{R}^{n\times n}$;

2) 构造矩阵 $\boldsymbol{D}$ 为度矩阵,度矩阵主对角线上的元素 $D(i,i)$ 为相似性矩阵 $\boldsymbol{S}$ 的第 i 行元素之和,构造拉普拉斯矩阵 $\boldsymbol{L}=\boldsymbol{D}^{-1/2}\boldsymbol{S}\boldsymbol{D}^{-1/2}$;

3) 拉普拉斯矩阵 $\boldsymbol{L}$ 进行特征分解,选取合适的特征向量按列存储构成矩阵 $\boldsymbol{Y}$;

4) 特征矩阵 $\boldsymbol{Y}$ 的每一列向量看着一个独立的样本,使用 k-means 对列向量进行聚类;

5) 最初的样本点 s_i 划分为第 j 聚类,当且仅当矩阵 $\boldsymbol{Y}$ 的第 i 列被划分为第 j 聚类 .

2.2 谱聚类在协同过滤算法中的应用

协同过滤算法主要分为基于用户的协同过滤(User-Based)和基于项目的协同过滤(Item-Based),本文主要考虑的是基于用户的协同过滤,在寻找相似的用户时,主要利用评分矩阵中用户对共同项的评分,常用 Pearson 相关系数度量其相似性. 一般来说,不同用户之间的共同评分数据较为稀疏,但评分数值可能比较接近,导致最后计算的相似性较为接近. 而这种情况可能是高估用户的相似性,如 2 个用户同时感兴趣的项目个数很少,那么他们的兴趣爱好可能很不相似.所以在使用 Pearson 相关系数计算用户相似度的同时,需要考虑用户的共同爱好的数量,因此本文使用一种修正的 Pearson 相关系数来定义用户的相似度.

$$r\rho_{u,v}=\frac{|N(u)\cap N(v)|}{|N(u)|\cup|N(v)|}\frac{\mathrm{cov}(u,v)}{\sigma_u\sigma_v}. \tag{6}$$

基于用户谱聚类的协同过滤推荐算法可分为用户信息聚类和内容推荐阶段,具体步骤:

用户信息聚类阶段:

步骤 1:获取用户的原始评分数据,构建评分矩阵 $\boldsymbol{S}$ 并对评分矩阵 $\boldsymbol{S}$ 进行数值归一化和平滑处理得到矩阵 $\boldsymbol{S}_{\text{norm}}$;

步骤 2:对于评分 $\boldsymbol{S}_{\text{norm}}$,使用式(6)计算各用户之间的相似度,得到用户相似度矩阵 $\boldsymbol{R}_{\text{sim}}$;

步骤 3:$\boldsymbol{R}_{\text{sim}}$ 作为谱聚类算法的输入,得出用户的聚类结果 $C(c_1,c_2,\cdots,c_k)$.

内容推荐阶段:

步骤 1:从待推荐用户所属归类中筛选出该用户的近邻用户,组成该用户的近邻集合 $\boldsymbol{N}(c_1,c_2,\cdots,c_n)$;

步骤 2:根据步骤 1 的结果,通过邻近集合 $\boldsymbol{N}(c_1,c_2,\cdots,c_n)$,并利用用户的相似度计算出用户未评分项目的预测评分值,生成用户预测评分向量 $\boldsymbol{P}$;

步骤 3:按照评分数值的高低对预测评分向量 $\boldsymbol{P}$ 进行排序,选择其中的 Top-N 作为推荐内容.

3　内容推荐与协同过滤的融合

在得到用户的内容推荐模型和协同过滤模型之后,为了保证推荐的准确性,兼顾 2 种推荐算法的效果,就需要融合 2 种模型的推荐结果.本文提出了一种基于逻辑回归的混合推荐模型,定义为

定义 3　融合推荐模型(Fusion User Recommended Model,FURM).对于任意用户,经过协同过滤,得到待推荐内容 $\boldsymbol{D}_{\text{r}}=\{\boldsymbol{d}_1,\boldsymbol{d}_2,\cdots,\boldsymbol{d}_n\}$($\boldsymbol{d}_i$ 为文本向量),及对应的预测评分 $\boldsymbol{W}_{\text{cf}}=\{w_1,w_2,\cdots,w_n\}$.$D_{\text{r}}$ 中的内容经过内容推荐算法处理得到相似度向量 $\boldsymbol{S}_{\text{cb}}=\{s_1,s_2,\cdots,s_n\}$;将 $\boldsymbol{W}_{\text{cf}}$ 与 $\boldsymbol{S}_{\text{cb}}$ 按照规则合并得到权值向量称为混合推荐模型,其表达式为 $\mathbf{FM}=\{\omega_1,\omega_2,\cdots,\omega_n\}$,其中 ω_i 为融合后的权值.

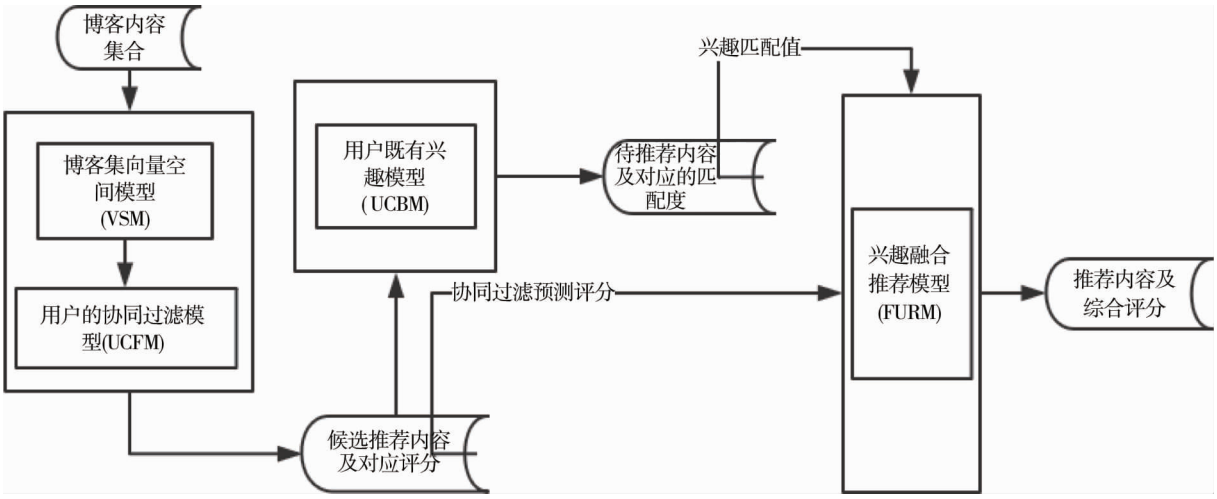

图 3　兴趣融合模型的工作流程图

采用基于逻辑回归的方法将 2 种推荐结果的值进行融合,W_{cf} 与 S_{cb} 的值作为式(7)的参数,参与运算.

$$W_\theta(X)=\frac{1}{1+\mathrm{e}^{-h_\theta(X)}}. \tag{7}$$

式中:$h_\theta(X)$ 可表示为

$$h_\theta(X)=\theta_0+\theta_1x_1+\theta_2x_2+\cdots+\theta_nx_n. \tag{8}$$

式中:$(\theta_0,\theta_1,\cdots,\theta_n)$ 为常数系数,通过大量的样本数据使用梯度下降的方法,对参数进行估计.

混合推荐模型的算法流程:

步骤 1:使用协同过滤算法获得待选推荐内容 $\boldsymbol{D}_{\text{r}}=\{d_1,d_2,\cdots,d_n\}$ 及其所对应的评分参数列表 $\boldsymbol{W}_{\text{cf}}=\{w_1,w_2,\cdots,w_n\}$;

步骤 2:将待选推荐内容 $\boldsymbol{D}_{\text{r}}=\{d_1,d_2,\cdots,d_n\}$ 作为参数输入到内容推荐模型,得到评分向量 $\boldsymbol{W}_{\text{cb}}=\{s_1,s_2,\cdots,s_n\}$;

步骤 3: $\boldsymbol{W}_{\mathrm{cf}}=\{w_1,w_2,\cdots,w_n\}$ 和 $\boldsymbol{W}_{\mathrm{cb}}=\{s_1,s_2,\cdots,s_n\}$ 作为式(7)的参数输入,得到最终的加权评分向量 $\boldsymbol{M}_{\mathrm{f}}=\{\omega_1,\omega_2,\cdots,\omega_n\}$;

步骤 4:对 $\boldsymbol{M}_{\mathrm{f}}=\{\omega_1,\omega_2,\cdots,\omega_n\}$ 进行排序,使与其对应内容列表 $\boldsymbol{D}_{\mathrm{r}}=\{d_1,d_2,\cdots,d_n\}$ 中的 Top-N 的内容作为最终的推荐内容.

算法 3 FURM 构建算法

输入 内容推荐模型:UCBM,协同过滤模型:UCFM,回归参数 $(\theta_0,\theta_1,\cdots,\theta_n)$ 及 $W_\theta(X)$;

输出 混合推荐模型 $\boldsymbol{M}_{\mathrm{f}}$.

Begin

Step1 通过 UCFM 协同过滤模型,获取待推荐的内容列表://内容及评分

UCFM $\rightarrow\{\boldsymbol{D}_r,\boldsymbol{W}_{\mathrm{cf}}\}$

Step2 $\boldsymbol{D}_r$ 作为内容推荐模型 UCBM 的输入,得到评分向量 $\boldsymbol{W}_{\mathrm{cb}}$:UCBM$(\boldsymbol{D}_{\mathrm{r}})\rightarrow\boldsymbol{W}_{\mathrm{cb}}$

Step3 $\boldsymbol{W}_{\mathrm{cb}}$ 和 W_{cf} 作为 $W_\theta(X)$ 的输入参数计算综合评分 ω_i, 并组成模型 **FM**;

Step4 对 **FM** 中的评分排序,输出 $\boldsymbol{D}_r$ 中对应的评分 Top-N 内容作为推荐结果.

End

4 实验设计与分析

4.1 实验数据集及评价指标

在实验数据集方面,本文的实验数据采用公开数据集 MovieLens 的 ml-latest-small (https://grouplens.org/datasets/movielens/)和 Book-Crossing.具体到 ml-latest-small,其中包含了来自 610 个用户对 9 742 部电影的 100 836 次评分;而 Book-Crossing 是对于阅读书籍的评分数据,其中包含 278 858 个用户对 271 379 本书的 1 149 780 次评分,由于该数据集数量较大,从中剔除评分次数少于 35 次的用户和书籍,最后得到的实验数据包含了 23 225 次评分.对于给定的实验数据集,随机抽取 20%的样本作为测试集,剩下的 80%作为训练集,采取交叉验证的方式.

本文选择准确率(Precision)和召回率(Recall)作为实验的评价指标.

$$\text{准确率(Precision)}=\frac{\text{用户点击的推荐数量}}{\text{推荐给用户的总数量}};\tag{9}$$

$$\text{召回率(Recall)}=\frac{\text{用户点击的推荐数量}}{\text{用户点击的总数量}}.\tag{10}$$

通过准确率衡量推荐算法的推荐内容与用户兴趣匹配的准确程度,召回率衡量算法发掘用户兴趣的范围的广度.

系统的平均绝对偏差(Mean Absolute Error,MAE)作为衡量系统的准确率的重要指标之一,是指推荐系统预测的用户对待推荐项的评分与用户实际的点击情况的误差状况,误差数值越小,说明推荐系统对用户喜好感知的越准确.

$$\mathrm{MAE}=\frac{1}{m}\sum_{i=1}^{m}|(y_i-\overline{y_i})|.\tag{11}$$

式中: y_i 为实际的用户点击概率且 y_i 取值为 $\{0,1\}$; $\overline{y_i}$ 为推荐系统预测的点击概率且 $\overline{y_i}\in[0,1]$.

4.2 实验设计及结果分析

实验以传统的基于内容的推荐算法、协同过滤算法为对照基准,本文提出的算法是基于协同过滤与内容推荐的融合算法,并以 3 种算法的结果进行对比分析(见表 1).

通过表 1 的数据可以看出,在待推荐的数据量较少的时候,3 种算法的 MAE 值都处于较低的水平,随着推荐数据的增加,MAE 值也伴随着增长,并且最终会趋于一个相对稳定的位置,总体来看,基于内容的推荐的误差较大,而本文提出的融合推荐模型 FURM 的 MAE 值普遍低于其他 2 种方法.为了直观的比较,给出 3 种算法的 MAE 值折线图如图 4 所示.

表 1　3 种算法平均绝对偏差值对比

推荐数目	基于内容的的推荐	协同过滤推荐	FURM 推荐算法
10	0.578	0.581	0.573
15	0.583	0.584	0.574
20	0.585	0.588	0.576
25	0.589	0.592	0.579
30	0.591	0.594	0.581
35	0.601	0.598	0.583
40	0.612	0.607	0.584
45	0.615	0.609	0.585
50	0.622	0.611	0.587
55	0.623	0.610	0.586
60	0.621	0.609	0.585

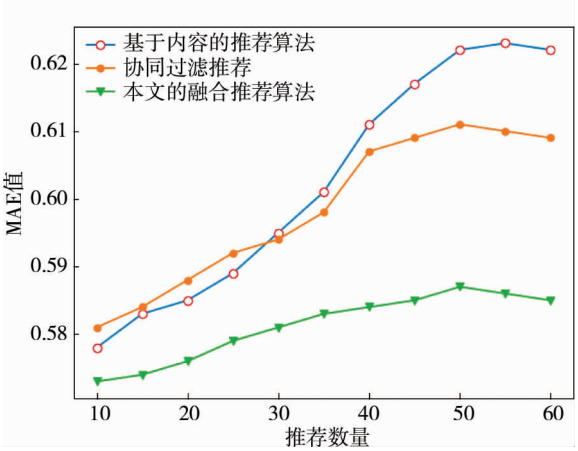

图 4　3 种算法的 MAE 值对比

选择推荐列表长度为 10~60 时进行算法召回率和准确率的实验,实验结果如表 2 所示.

表 2　3 种算法准确率和召回率对比

推荐数目	基于内容的推荐		协同过滤推荐		FURM 推荐算法	
	准确率	召回率	准确率	召回率	准确率	召回率
10	0.814	0.574	0.783	0.576	0.801	0.575
15	0.796	0.577	0.781	0.577	0.793	0.576
20	0.794	0.583	0.776	0.581	0.782	0.577
25	0.787	0.586	0.771	0.588	0.781	0.581
30	0.763	0.591	0.765	0.590	0.775	0.589
35	0.761	0.593	0.761	0.593	0.776	0.591
40	0.759	0.585	0.757	0.596	0.769	0.594
45	0.755	0.581	0.753	0.592	0.767	0.589
50	0.757	0.576	0.754	0.591	0.768	0.585
55	0.756	0.572	0.756	0.594	0.764	0.584
60	0.754	0.573	0.759	0.593	0.766	0.587

从算法的准确率来看,3 种算法的准确率都保持在较高的水平,随着推荐数量的增多,呈现出较为明显的下降趋势,在推荐数量较少的情况下,基于内容的推荐算法的准确率较其他 2 种算法较高,随后出现一定的下降趋势,在推荐数量较多的情况下,协同过滤算法会有一定的优势,除此之外,本文所提出的算法在准确率方面表现的比较平稳.

通过表 2 可以看出,3 种算法的召回率都处于较好的水平.比较可知,基于内容的推荐算法,在做较大数量的推荐时其召回率不高,主要是传统的基于内容推荐算法对用户兴趣的广度发掘不够;而本文提出的算法与协同过滤算法类似在召回率上表现的较好,比较适合较大规模的数据推荐.如图 6 所示,给出 3 种算法在召回率上的对比关系.

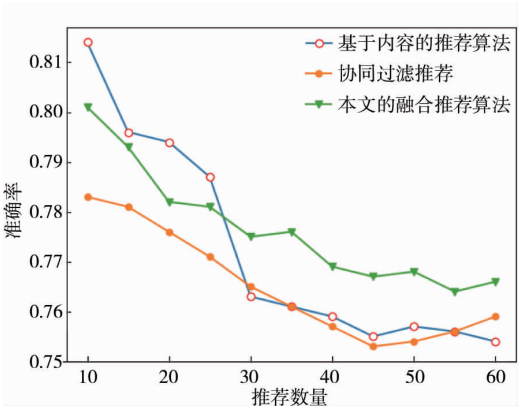

图 5　3 种算法的准确率对比

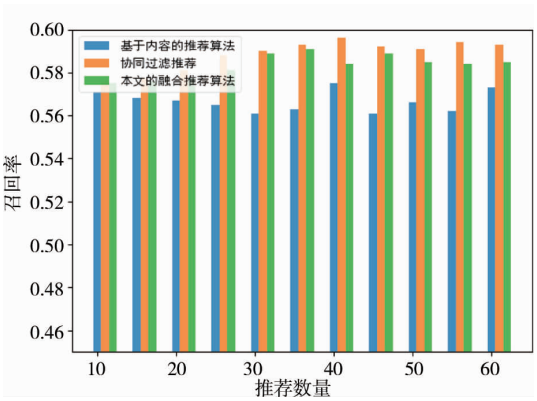

图 6　3 种算法的召回率比较

5 结论

1) 提出了一种基于内容和协同过滤的混合推荐方案,将时间因素对用户兴趣的影响引入到推荐模型的构建过程中,并进一步利用用户的反馈机制对最终推荐结果进行过滤.

2) 本文所提的混合推荐方案相较于基于内容的推荐算法和基于协同过滤的算法,在准确率和召回率上都有明显的提升,且该混合方案在多个应用场景(如书籍,电影等)中都具有较好的推荐效果.后续研究将考虑如何进一步提高融合模型的效率和准确度.

参考文献:

[1] Langford J, Li L, Zhang T. Sparse online learning via truncated gradient[J]. Journal of Machine Learning Research, 2008, 10(2): 777-801.

[2] 文凯, 朱传亮. 融合社交网络和兴趣的正则化矩阵分解推荐模型[J]. 计算机应用, 2018, 38(9): 2523-2528.

[3] Tsi C F, Hung C. Cluster ensembles in collaborative filtering recommendation[J]. Applied Soft Computing, 2012, 12(4): 1417-1425.

[4] 杨琛, 李秉智. MPLS 多播机制中模糊标签聚集的研究[J]. 计算机工程与应用, 2007, 43(36): 147-149.

[5] Panigrahi S, Lenka R K, Stitipragyan A. A hybrid distributed collaborative filtering recommender engine using apache spark [J]. Procedia Computer Science, 2016, 83(1): 1000-1006.

[6] 付永平,邱玉辉.一种基于贝叶斯网络的个性化协同过滤推荐方法研究[J].计算机科学,2016,43(9):266-268.

[7] Liao C L, Lee S J. A clustering based approach to improving the efficiency of collaborative filtering recommendation[J]. Electronic Commerce Research & Applications, 2016, 18: 1-9.

[8] Maulik U, Chakraborty D. A self-trained ensemble with semi supervised SVM:an application to pixel classification of remote sensing imagery[J]. Pattern Recognition, 2011, 44(3): 615-623.

[9] Liu K, Guo Y, Wang S, et al. Semi-supervised learning based on improved co-training by committee[C]// International Conference on Intelligence Science and Big Data Engineering, 2015: 413-421.

[10] Yu L , Liu L , LI X F. A hybrid collaborative filtering method for multiple interests and multiple content recommendation in E-Commerce[J]. Expert Systems with Applications ,2005 ,28 (1): 67-77.

[11] Cheng S, Liu Y. Time-aware and grey incidence theory based user interest modeling for document recommendation[J]. Cybernetics and Information Technologies, 2015, 15(2): 36-52.

[12] Quijano-Sánchez L, Díaz-Agudo B, Recio-García J A. Development of a group recommender application in a Social Network [J]. Knowledge-Based Systems, 2014, 71: 72-85.

[13] 魏慧娟,戴牡红,宁勇余. 基于最近邻居聚类的协同过滤推荐算法[J].中国科学技术大学学报,2016(9):736-742.

[14] 刘丽芳.微博客的传播特征与传播效果研究[D].杭州:浙江大学,2010.

[15] 王媛媛,李翔. 基于人口统计学的改进聚类模型协同过滤算法[J].计算机科学,2017,44(3):63-69.

[16] 罗斌,陈翔.幂律特性在新浪微博个性化推荐中的应用研究[J].计算机工程与科学,2018,40(4):731-739..

[17] Zhao S, Yang X, Li X, et al. A Hadoop-based visualization and diagnosis framework for earth science data[C]//Proceedings of the2015 IEEE International Conference on Big Data. Piscataway. NJ:IEEE, 2015: 1972-1977.

[18] 周林,平西建,徐森,等.基于谱聚类的聚类集成算法[J].自动化学报,2012,38(8):1335-1342.

[19] 赵小强,刘晓丽.基于密度敏感的改进自适应谱聚类算法[J].兰州理工大学学报,2018,44(6):102-106.

[20] 杨随心,耿修瑞,杨炜暾,等.一种基于谱聚类算法的高光谱遥感图像分类方法[J].中国科学院大学学报,2019(2):267-274.