doi: 10.3969/j.issn.0490-6756.2018.01.009

基于改进协同过滤算法的个性化新闻推荐技术

黄贤英, 熊李媛, 李沁东
(重庆理工大学计算机科学与工程学院, 重庆 400054)

摘 要: 针对传统的基于内容协同过滤算法只是依据用户历史访问矩阵向用户做出推荐,存在数据稀疏以及不能及时反映用户兴趣变化等问题,个性化新闻推荐技术在传统的协同过滤算法基础上提出了新闻文本内容相似度的计算方式和时间窗的概念,新闻内容相似度计算中还考虑了特征词的词性和在新闻中的位置的影响,时间窗用来建立适应用户兴趣随时间变化的模型;实验结果表明,改进后的算法有效地改善了新闻用户历史访问数据的稀疏问题,及时捕获用户兴趣,F-measure 值相比传统的算法最大提高了 11.5%,平均绝对误差值最高下降了 8%,显著提高了推荐质量.

关键词: 新闻推荐; 协同过滤; 内容相似度; 时间窗
中图分类号: TP301.6 **文献标识码**: A **文章编号**: 0490-6756(2018)01-0049-07

Personalized news recommendation technology based on improved collaborative filtering algorithm

HUANG Xian-Ying, XIONG Li-Yuan, LI Qin-Dong
(College of Computer Science and Engineering, Chongqing University of Technology, Chongqing 400054, China)

Abstract: The traditional collaborative filtering algorithm only based on matrix produced by user access history to make recommendation and sparse data, and also cannot reflect the user's interests timely, contrary to these problems, the personalized recommendation technology news in the traditional collaborative filtering algorithm proposes the calculation of news text content similarity and the concept of the time window, the calculation of news content similarity also takes into account the part of speech and positions of the feature words in the news, the time window is used to create user interest model which will change over time; The experimental results show that the improved algorithm effectively improves the sparse problem of data which user has accessed and captures user interest timely, F-measure value improves the maximum 11.5% compared to the traditional algorithm, the highest value of mean absolute error fell by 8%, greatly improving the quality of recommendation.

Keywords: News recommendation; Collaborative filtering; Connect similarity; Time window

1 引 言

由于互联网技术的飞速发展,各大新闻网站的迅猛增加,新闻信息量每天呈指数增长.用户阅读自己感兴趣的新闻成了难题,个性化推荐技术便应运而生.个性化新闻推荐技术根据不同新闻用户的历史浏览行为,自动、高效地给用户推荐新闻,这一技术得到了用户的广泛认可.

收稿日期: 2016-09-23
基金项目: 教育部人文社科青年基金项目(16YJC860010); 重庆市社会科学规划项目(2015BS059); 国家自然科学基金项目(61603065)
作者简介: 黄贤英(1967-), 女, 重庆人, 教授, 硕士, 研究方向为信息检索、移动计算. E-mail: 1303366922@qq.com
通讯作者: 熊李媛. E-mail: 623890251@qq.com

协同过滤推荐算法是目前为止应用较为广泛的个性化推荐技术.传统的协同过滤算法存在一些问题:首先,用户历史访问矩阵的稀疏问题会导致相似度的计算结果不够准确.另外,传统算法中没考虑用户的兴趣随时间改变的问题.针对这些问题,国内外学者进行了很多研究.沈西挺等[1]考虑到用户间的相似性不仅与用户评过分的项目有关,还与用户本身对项目的感兴趣程度有关,提出了基于用户兴趣的相似性计算方式,这一想法一定程度上减少了传统算法中数据稀疏带来的负面影响.邢春晓等[2]从用户兴趣会发生衰减和反复性变化的角度出发,提出了基于时间和基于资源相似度的数据权重,构建了适应用户兴趣变化的用户模型,蒋崇礼等[3]在此基础上考虑了项目属性的影响,这对于数据稀疏问题有一定的缓解.杨武[4],史艳翠[5]都根据用户的评价矩阵和带时间权的新闻特征词矩阵计算用户间的混合相似度值,有效地改善了传统算法中的数据稀疏问题.赖雯[6]在此基础上增加了项目聚类的算法,根据项目的属性采用 K-means 算法实现聚类,但对于新闻来讲,只根据属性聚类是不够准确的,因为新闻中的热点词对于新闻有较大的影响.文献[7,8]在协同过滤算法的基础上增加了皮尔逊相关系数,提高了计算用户间的相关性准确率.文献[9,10]将基于项目的协同过滤算法和基于用户群的人口统计相结合,这有效解决了用户的冷启动问题.

本文针对新闻内容本身和用户兴趣的改变对新闻推荐都有重要影响,提出了融合新闻内容和用户兴趣的改进的协同过滤算法(Improved Collaborative Filtering Algorithm of Integrating News Content and User Interests, ICFANU)来对新闻进行个性化推荐.在传统的协同过滤算法的基础上根据新闻文本的特点改进了新闻相似度的计算方式,同时考虑了新闻访问相似度和新闻内容相似度,新闻内容相似度根据词语的词性和在新闻中所处的位置来计算;传统的协同过滤算法没有考虑用户兴趣随时间的变化,引入时间窗概念,建立用户近期兴趣模型和用户行为反复的兴趣模型,最终得到用户混合兴趣模型,在 DataCastle 提供的新闻数据集进行实验,实验研究证明,本文算法在推荐结果上相对传统的协同过滤算法和已有改进过的算法都有显著提高.

2 传统的协同过滤算法

传统的基于项目的协同过滤算法中,用户访问新闻的历史数据可以表示成矩阵 D_{n*m} 的形式.

$$D_{n*m}=\begin{bmatrix}D_1\\D_2\\\vdots\\D_N\end{bmatrix}=\begin{bmatrix}d_{11}&d_{12}&\cdots&d_{1m}\\d_{21}&d_{22}&\cdots&d_{2m}\\\vdots&\vdots&\vdots&\vdots\\d_{n1}&d_{n2}&\cdots&d_{mn}\end{bmatrix} \tag{1}$$

其中,n 表示访问新闻系统的用户总数;m 表示 n 个用户访问过的新闻总数.d_{ij} 代表用户 i 是否阅读过新闻 j,1 代表阅读,0 代表未阅读.

用 $x(j)$ 表示阅读新闻 j 的用户数,其中,$x(j)=\sum_{m\in n}v_{mj}$.根据矩阵 D 建立新闻-用户的倒排表,倒排表形式如表 1 所示.

表 1 新闻-用户访问倒排表

Tab. 1 The inverted table to access of news-users

	1	2	…	i	…	m
1	x_{11}	x_{12}	…	x_{1i}	…	x_{1m}
…	…	…	…	…	…	…
j	x_{j1}	x_{j2}	…	x_{ji}	…	x_{jm}
…	…	…	…	…	…	…
m	x_{m1}	x_{m2}	…	x_{mi}	…	x_{mn}

其中 1,2,3…,m 表示 n 个用户阅读过的新闻,x_{ij} 代表同时阅读新闻 i 和新闻 j 的用户数,因此 $x_{ij}=x_{ji}$.新闻 i 和新闻 j 的访问相似度 sim_1 计算公式如式(2)所示.

$$\mathrm{sim}_1(i,j)=x_{ij}/[x(i)*x(j)] \tag{2}$$

传统的协同过滤算法(Item-CF)核心是根据式(2),计算用户访问过的新闻集内的每条新闻和欲推荐新闻的访问相似度,得到相似度较高的新闻推荐给用户.Item-CF 算法的流程如下.

算法 1 Item-CF **算法**

输入 用户 u 的历史访问数据集 D_u,欲推荐的新闻集 M

输出 用户 u 的 top-N 推荐

1) 对于用户 u 阅读过的每条新闻 $i\in D_u$,通过公式(1)、欲推荐的新闻集 M,计算新闻 i 和欲推荐的每条新闻 $j\in M$ 的相似度,取相似度较高的新闻作为新闻 i 的邻居集 C_i,合并用户 u 的所有邻居集并删除已存在的新闻得到用户 u 的候选集 C,其中 $C=\sum_{i\in D_u}C_i$;

2) 计算候选集 C 中的新闻 j 与用户 u 中每条新闻的相似度之和,如公式(3)所示.

$$P_{uj}=\sum_{i\in D_u}[\mathrm{sim}_1(i,j)*d_{ui}] \tag{3}$$

其中,d_{ui} 表示用户 u 对于新闻 i 的兴趣,若用户对

于新闻 i 有过访问记录，则 $d_{ui}=1$，否则 $d_{ui}=0$.

3) 根据 P_{uj} 值从大到小进行排序，取前 N 条新闻推荐给用户 u.

传统的 Item-CF 算法虽然得到了广泛应用，但是作为新闻推荐没有考虑新闻内容本身对新闻推荐的重要性、用户的兴趣会随时间发生改变的问题，因此，还需改进协同过滤算法中相似度的计算方式，建立适合用户兴趣变化的模型.

3　改进的新闻相似度计算方式

由于新闻的文本特性，新闻的相似度不仅需要考虑基于用户访问的相似度，还要考虑新闻内容的相似性. 另外，在新闻文本中，不同词性的词语在文本中的重要性不相同、不同位置的词语的重要性不同，因此，在计算新闻内容相似度的同时还需考虑词语的词性和位置的影响.

3.1　新闻特点

1) 新闻标题中的词语重要性高于新闻内容中词语的重要性. 用户浏览一篇新闻时，首先看到的是新闻的标题，标题符合用户的兴趣时，用户才打开这篇新闻进行浏览，因此，对于用户来说，新闻标题中词语的重要性高于新闻内容中词语的重要性.

2) 新闻文本中不同词性的词语重要性不同. 新闻的重要载体是文本内容，根据文本信息的特点，名词或者动词在文章中的重要性一般高于其他词性的词语. 因此，计算新闻之间相似度时可以根据不同词性计算词语的权重.

3.2　改进的新闻相似度的计算

传统的协同过滤只是基于用户共同访问相同新闻数得到新闻相似度，没有考虑新闻内容的相似度. 由于新闻内容是文本型的，可以提取特征词采用向量空间模型来表示. 首先，对于文章中出现的符号和定义做出说明.

定义 1　特征词. 对于给定的某篇新闻，能够准确、合理地代表或者描述此新闻内容的单词或者成语叫做该新闻的特征词 f. 由特征词构成的特征序列 $F=\{f_1,f_2,\cdots,f_k\}$ 称为此篇新闻的特征词序列. 其中 k 代表特征词的个数.

根据新闻特点 1)，在计算新闻中特征词的权重时，标题中的特征词权重要高于正文中特征词的权重，位置权重因子的表示如公式(4)所示.

$$\begin{cases}\vartheta, & \text{特征词出现在标题中}\\ \sigma, & \text{特征词出现在正文中}\end{cases} \tag{4}$$

其中，$0\leqslant\sigma\leqslant\vartheta\leqslant1$，在新闻中特征词的位置权重因子的计算方式如(5)所示.

$$pt_{ij}=\frac{freq'(i,j)*\vartheta+freq''(i,j)*\sigma}{freq(i,j)} \tag{5}$$

其中，$freq'(i,j)$ 表示特征词在新闻的标题中出现的次数，$freq''(i,j)$ 表示特征词在新闻的正文中出现的次数.

根据新闻的文本特点 2)，新闻的表示方式可以采用向量空间模型，对于给定的新闻和特征词序列 $F=\{f_1,f_2,\cdots,f_k\}$ 可以得到新闻的向量空间模型表示为 $d_i=(w_1,w_2,\cdots,w_k)$，w_i 代表对应特征词 f_i 的权重，即该词在文档中的重要程度值. 对于整个新闻集的向量空间模型表示如式(6)所示.

$$DW_{m*k}=\begin{bmatrix}w_{11} & w_{12} & \cdots & w_{1k}\\ w_{21} & w_{22} & \cdots & w_{2k}\\ \vdots & \vdots & \vdots & \vdots\\ w_{m1} & w_{m2} & \cdots & w_{mk}\end{bmatrix} \tag{6}$$

其中，权重的计算方式采用的 $tf\text{-}idf$ 计算法，w_{ij} 表示第 i 篇新闻中的第 j 个词语的权重.

在权重的计算中加入词性系数. 词性系数表示如公式(7)所示.

$$\theta=\begin{cases}\alpha, & f=n,f=v\\ \beta, & f=adj,f=adv\\ 0, & \text{other}\end{cases} \tag{7}$$

其中，$0\leqslant\beta\leqslant1$. 因此特征词 j 在新闻中的权重计算方式如(8)所示.

$$w'_{ij}=w_{ij}*\theta*pt_{ij} \tag{8}$$

新闻集中的每条新闻最终的向量空间模型表示如(9)所示.

$$DW'_{m*k}=\begin{bmatrix}w'_{11} & w'_{12} & \cdots & w'_{1k}\\ w'_{21} & w'_{22} & \cdots & w'_{2k}\\ \vdots & \vdots & \vdots & \vdots\\ w'_{m1} & w'_{m2} & \cdots & w'_{mk}\end{bmatrix} \tag{9}$$

新闻内容相似度计算方式可以采用内积的方法、余弦相似度计算方法等. 此处采用的是余弦相似度计算方法，计算公式如(10)所示.

$$\mathrm{sim}_2(i,j)=\frac{DW_i'*DW_j'}{|DW_i'|*|DW_j'|} \tag{10}$$

其中 DW_i' 代表新闻 i 的向量空间模型，DW_j' 代表新闻 j 的向量空间模型.

公式(2)中的新闻相似度计算方式是从用户阅读行为考虑，公式(9)中的计算方式是从新闻文本特点考虑，两者的结合可以得到新闻相似度混合计算公式为式(11).

$$\mathrm{sim}(i,j)=\mu\mathrm{sim}_1(i,j)+(1-\mu)\mathrm{sim}_2(i,j) \tag{11}$$

其中，$0\leqslant\mu\leqslant1$.

4 用户兴趣模型的建立

用户的兴趣会随着时间发生改变,一般来说,用户近期的浏览行为反映的是用户最近的兴趣,对用户的推荐结果有较重要的影响,而早期的浏览行为对用户的推荐结果的影响较小. 另外,用户的兴趣是周期性的,这会造成用户早期的浏览行为对用户的推荐也有重要影响. 因此提出时间窗,来获取用户稳定的兴趣模型.

4.1 时间窗

时间窗就是把用户曾经浏览新闻的一段时间作为用户的兴趣时间段,用来对用户阅读过的其他新闻进行加权. 时间窗的选取一般是把用户最近浏览新闻的时间作为一个时间点,向前再选取一个时间点,这两个时间点之间的间隔就是一个时间窗. 这样选取的原因是用户近期浏览的新闻一般代表用户的近期兴趣.

通过时间窗得到用户 u 在 T 时间段内浏览过的新闻集合 D_{uT},对于在时间窗之外用户浏览过的新闻 $i \in D_u$,不管访问时间早晚,如果和 D_{uT} 内新闻有很高的相似度,则说明其属于用户 u 的兴趣. 那么,对于用户 u 产生推荐也有重要的影响. 时间窗 T 的大小需要根据实验结果来确定.

4.2 模型的建立

根据时间窗的概念,用户的兴趣模型要从两方面考虑.

1) 基于用户近期行为的兴趣模型

文献[11]中提出了人的遗忘规律:人的遗忘速度会随着时间变得越来越慢. 并且将这种规律用到了推荐算法中以此来表示用户兴趣的衰减,时间权重度量公式如公式(12)所示.

$$WR(u,i) = e^{-\varepsilon \times \frac{Date_{interval}(i)}{L_u}} \tag{12}$$

其中,L_u 表示用户 u 使用推荐系统的时间跨度,$Date_{\text{interval}}(i)$表示用户对新闻的浏览时间与用户对最新一条新闻的浏览时间的间隔. $0 < \varepsilon < 1$,适合推荐系统的值通过实验得到.

2) 基于用户行为反复的兴趣模型

公式(12)中的时间度量公式只能建立基于用户近期行为的兴趣模型. 根据时间窗概念,还需建立基于用户行为反复的兴趣模型.

首先,根据用户 u 的已访问新闻集 D_u,定义一个时间窗 T,然后获取用户 u 在最近 T 时间段内浏览过的新闻集 D_{uT}. 用户 u 早期阅读过的新闻 i 和用户 u 在 T 时间段内新闻相似度的计算采用平均相似度,即新闻 i 和新闻集 D_{uT} 中每条新闻的相似度值和的平均值,当达到一定的相似度值就认为新闻和用户 u 的近期兴趣相似. 具体公式如(13)所示.

$$WN(u,i) = \overline{\text{sim}(i,DuT)} = \frac{\sum_{j \in DuT} \text{sim}(i,j)}{\text{size}(DuT)} \tag{13}$$

其中,sim(i,j)的计算方式采用的是公式(11),size(D_{uT})表示用户 u 在最近 T 时间段内阅读新闻的总条目数.

3) 两种模型的融合

通过以上两种模型的建立可以看出:每种模型的侧重点不同. 前者主要考虑用户的近期行为与用户的兴趣更为相似,后者考虑用户兴趣的反复性,即用户早期的兴趣也许对于用户的兴趣有更为重要的影响. 因此,在建立用户的兴趣模型时要综合考虑两方面的影响. 两种模型的融合采用的是按照一定的比例因子进行相加. 具体公式如(14)所示.

$$WT(u,i) = \varphi WR(u,i) + (1-\varphi) WN(u,i) \tag{14}$$

其中,$0 \leqslant \varphi \leqslant 1$,$\varphi$ 和 $(1-\varphi)$ 分别代表两种模型所占的比例,具体 φ 值通过实验结果来确定.

5 基于改进的协同过滤算法的个性化推荐过程

通过上面的改进,个性化新闻推荐技术不仅考虑了新闻内容的相似度计算,还建立了基于用户兴趣变化的模型,具体的推荐过程如算法 2.

算法 2 基于用户兴趣变化的模型

输入 用户 u 的历史访问数据集 D_u,倒排表 1,欲推荐的新闻集 M

输出 用户 u 的 top-N 推荐

1) 根据用户 u 的历史访问数据集 D_u、公式(11),(14)得到用户 u 的混合兴趣模型,即得到用户 u 感兴趣的新闻集 D_{ui}.

2) 对于用户 u 感兴趣的每条新闻 $i \in D_{ui}$,通过表 1、欲推荐的新闻集 M,根据公式(11)计算新闻 i 和欲推荐的条新闻 $j \in M$ 的相似度 $sim(i,j)$,按照相似度的大小得到新闻 i 最近邻居集 C_i,合并用户 u 阅读过的每条新闻的邻居集,并删除临近集中已存在的新闻,得到用户 u 的新闻候选集 C,其中 $C = U_{i \in D_u} C_i$;

3) 计算候选集中的每条新闻 $j \in C$ 与用户感兴趣的新闻集中每条新闻 $i \in D_{ui}$ 推荐值之和,具体

公式如(15)所示.

$$\mathrm{rec}(u,j) = \sum_{i\in D_{ui}} [WT(u,i) * \mathrm{sim}(i,j)] \quad (15)$$

根据 rec(u,j)值从大到小进行排序，取前 N 条新闻推荐给用户 u.

6　实验结果及分析

6.1　数据集

此次实验采用的数据集是由 DataCastle 提供的，来自财新网上随机抽取的 10000 个用户在 2014 年 3 月的所有新闻浏览记录，每条新闻包括用户编号、新闻编号、浏览时间(精确到秒)以及新闻文本内容. 这些数据是 10000 个用户随机浏览新闻产生的，具有真实性、可靠性，能够保证实验结果的准确性.

这次实验 2 随机抽取三组，每组包括阅读新闻条目数超过 50 条，使用推荐系统跨度达到 20 天以上时间的用户各 200 名，其中，每个用户最后浏览的 20 条新闻作为测试集，其余浏览的新闻作为训练集，这样可以确保用户是长期、稳定使用新闻系统，使用推荐算法向这些用户进行推荐可以保证推荐结果有更高的准确性. 由于每组实验都会产生误差，误差的大小不确定，采用三组实验结果的平均值作为最终结果，可以减小误差.

6.2　评价标准

实验结果是根据每个用户在训练集中的访问记录向该用户推荐 N 条新闻，因此本次实验可以采用平均绝对误差(Mean Absolute Error，MAE)、F-measure 作为评价标准. F-measure 的评价公式如(16)所示.

$$F\text{-}measure = \frac{2}{1/P + 1/R} \quad (16)$$

其中，P、R 分别代表的是推荐结果的准确率和召回率，具体计算公式为 $P = \sum_{u_i\in U} hit(u_i) / \sum_{u_i\in U} L(u_i)$、$R = \sum_{u_i\in U} hit(u_i) / \sum_{u_i\in U} T(u_i)$. U 为数据集中 200 个用户的集合，$hit(u_i)$表示推荐给用户 u_i 的新闻中，真正在测试集中被该用户浏览的个数. $L(u_i)$表示提供给用户 u_i 的新闻推荐列表的长度，$T(u_i)$则为测试集中用户 u_i 真正浏览的新闻的数目，在本次实验中 $T(u_i) = 20$，$\sum_{u\in U} T(u_i) = 1000$.

MAE 是最常用的度量推荐系统质量好坏的标准，它是指评分偏差的绝对值的平均，就是在推荐系统中预测评分和实际评分之间的差异. 在此次推荐算法中，推荐结果并没有评分，只有是否推荐正确. 因此可以设置预测推荐结果是 1 和 0，1 代表推荐结果和用户实际浏览的新闻是一致的，相反不一致就是 0. MAE 的计算公式如(17)所示.

$$MAE = \frac{\sum_{i\in \mathbf{N}} |P_{u,i} - r_{u,i}|}{N} \quad (17)$$

其中，N 代表推荐的新闻条数；$P_{u,i}$ 代表用户浏览过的新闻，此处 $P_{u,i} = 1$，$r_{u,i}$ 代表系统推荐的新闻是否为用户浏览过，如果和用户的浏览过的新闻一致，则 $r_{u,i} = 1$，不一致则 $r_{u,i} = 0$，因此，MAE 值越小代表推荐算法的质量越高，反之，MAE 值越大则推荐质量越低.

6.3　实验结果

1) ICFANU 算法中参数的确定

首先，确定用户的临近集个数和新闻推荐的条数，为了更好的区分实验的结果，临近集个数为 30，推荐新闻数目设为 15. 其次，计算公式中的因子值，采用待定系数法，在三组训练集下分别计算 ICFANC 算法的 F-measure 值，再计算三者平均值，在平均值最大时候确定因子的值. 根据实验，首先确定了 $\mu = 0.9$，$\varepsilon = 0.1$，$\varphi = 0.5$ 时，F-measure 值最高.

在 μ，ε，φ 确定的情况下，调整公式(13)中 T 的大小，令 T 为 5、10、15、20、25、30，由图 1 可以看出当 $T = 25$ 时，F-measure 值最高，因此设公式(13)中的 T 为 25.

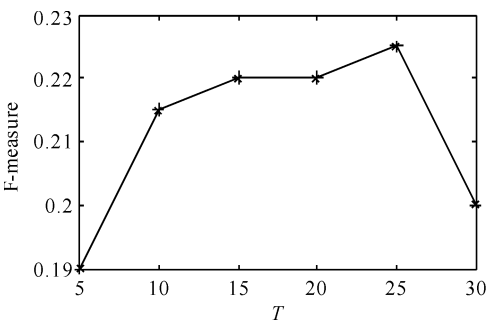

图 1　不同的 T 值对 ICFANU 算法的影响

Fig. 1　The influence of different T values on ICFANU algorithm

在 μ，ε，φ，T 确定的情况下，调整公式(7)中 α，β 因子的大小，由于在文本中名词或者动词的词性重要性最高，则认为 $\alpha = 1$、只改变 β 的值就可以调整动词或者名词与其他词汇的重要性差别，令 β 为 0，0.1，0.2，…，1，由图 2 可以看出当 $\beta = 0.1$ 时，F-measure 值最高，因此设公式(7)中的 β 为 0.1.

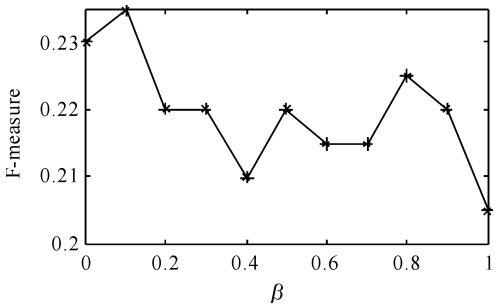

图 2 不同的 β 值对 ICFANU 算法的影响

Fig. 2 The influence of different β values on ICFANU algorithm

在 $\mu,\varepsilon,\varphi,T,\beta$ 的情况下，调整公式(4)中 ϑ,σ 因子的大小，由于在新闻标题中的词语重要性不低于在文本中词语的重要性，则可以令 $\vartheta=1$、只改变 σ 的值就可以调整新闻标题中的词语重要性与新闻文本中词语的重要性差别，令 σ 为 0，0.1，0.2，…，1，由图 3 可以看出当 $\sigma=0.4$，F-measure 值最高，因此设公式(4)中的 σ 为 0.4.

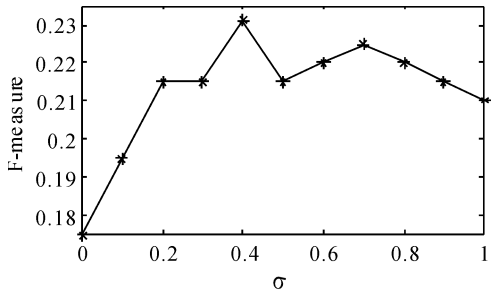

图 3 不同的 σ 值对 ICFANU 算法的影响

Fig. 3 The influence of different σ values on ICFANU algorithm

2) 对比实验

图 4 显示了在推荐新闻条数分别为 10，20，30 的情况下，本文中的推荐算法、传统的基于协同过滤算法、文献[3]、文献[6]算法的推荐结果的 F-measure 值，结果显示，本文的算法 F-measure 值最高，在推荐新闻条数为 20 时，F-measure 值达到最高.

图 5 显示了在新闻邻居集个数分别为 20，30，40，50，推荐新闻条目为 20 条的情况下，本文中的推荐算法、传统的协同过滤算法、文献[3]、文献[6]算法的推荐结果的 MAE 值. 结果显示，本文算法的 MAE 值在不同的邻居集个数下都能达到最小，说明本文算法推荐的差异值最小，推荐质量最高.

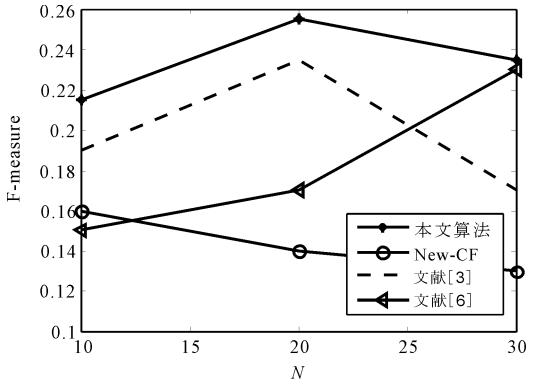

图 4 推荐条目不同情况下各算法的 F-measure 值

Fig. 4 The F-measure values for each algorithm in different recommended items situations

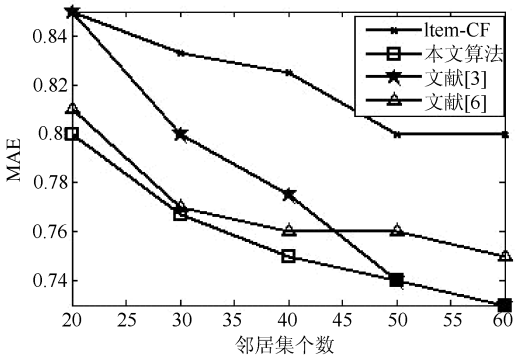

图 5 邻居集个数不同情况下各算法的 MAE 值

Fig. 5 The MAE values for each algorithm in different neighbor set situations

7 结 论

本文提出的基于改进的协同过滤推荐算法的个性化新闻推荐技术是在传统的协同过滤基础上同时考虑了访问相似度和内容相似度，在计算内容相似度的过程中考虑了不同词性的词语重要性不同和新闻标题中的词语重要性高于文本内容中的词语两个因素，这提高了新闻文本相似度计算的准确性，并有效缓解了数据集的稀疏问题；其次，根据时间窗，建立适应用户兴趣变化的模型，最终优化了推荐结果. 本文算法不仅比传统的协同过滤算法有很大提高，和已有的改进算法相比也有明显提高，相比蒋崇礼[3]提出的混合推荐算法 F-measure 值最大提高 11.5%，MAE 值最大下降 8%，相比赖雯[6]提出的算法 F-measure 值最大提高 1.5%，MAE 值最大下降 1%，这证明了本文算法有较好的推荐质量. 下一步，将考虑文本中语义之间的相似性以及聚类在新闻推荐中的影响，来进一步提高新闻推荐性能.

参考文献：

[1] 沈西挺，董智佳. 反应用户兴趣变化的协同过滤算法[J]. 计算机应用与软件，2013，30：295.

[2] 邢春晓，高凤荣，战思南，等. 适应用户兴趣变化的协同过滤推荐算法[J]. 计算机研究与发展，2007，44：296.

[3] 蒋崇礼，汪瑜彬. 一种个性化协同过滤混合推荐算法[J]. 软件导刊，2016，15：53.

[4] 杨武，唐瑞，卢玲，等. 基于内容的推荐与协同过滤融合的新闻推荐方法[J]. 计算机应用，2016，36：414.

[5] 史艳翠，戴浩男，石和平，等. 一种基于时间戳的新闻推荐模型[J]. 计算机应用与软件，2016，33(6)：40.

[6] 赖雯. 协同过滤推荐系统的用户兴趣变化和稀疏性问题研究[D]. 广州：华南理工大学，2013.

[7] Ge F. A collaborative filtering recommendation approach based on user rating similarity and user attribute similarity[J]. Adv Mat Res，2013，846-847：1736.

[8] Xia J X，Wu F，Xie C S. A novel similarity measure based on weighted bipartite network for collaborative filtering recommendation[J]. Appl Mech Mat，2013，263：1834.

[9] Vizine Pereira A L，Hruschka E R. Simultaneous co-clustering and learning to address the cold start problem in recommender systems[J]. Knowledge-Based Systems，2015，82(C)：11.

[10] Vozalis M G，Margaritis K G. Using SVD and demographic data for the enhancement of generalized Collaborative Filtering [J]. Inf Sci Internat J，2007，177：3017.

[11] 彭石，周志彬，王国军. 基于评分矩阵欲填充的协同过滤算法[J]. 计算机工程，2013，39：175.

[12] 刘汉清，朱敏，苏亚博，等. 一种考虑用户兴趣转移特征的协同预测模型[J]. 四川大学学报：自然科学版，2016，53：548.

[13] Bokde D K，Girase S，Mukhopadhyay D. An item-based collaborative filtering using dimensionality reduction techniques on mahout framework[J]. Comput Sci，2015，134：561.