计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2018,27(5):193−197 [doi: 10.15888/j.cnki.csa.006356]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于新闻时效性的协同过滤推荐算法①

冯文杰, 熊　翱

(北京邮电大学 网络技术研究院, 北京 100876)
通讯作者: 冯文杰, E-mail: tmacfeng@yeah.net

摘　要: 提出一种基于新闻时效性的协同过滤推荐算法. 首先对新闻的时效性进行了特征分析, 建立了新闻时效性模型, 然后结合新闻时效性改进了基于用户的协同过滤算法. 最后进行了仿真实验, 实验结果表明, 该方法可以有效提高推荐算法的性能, 改善新闻推荐准确度和召回率.

关键词: 协同过滤; 新闻推荐; 时效性模型

引用格式: 冯文杰,熊翱.基于新闻时效性的协同过滤推荐算法.计算机系统应用,2018,27(5):193–197. http://www.c-s-a.org.cn/1003-3254/6356.html

Collaborative Filtering Recommendation Algorithm Based on News Timeliness

FENG Wen-Jie, XIONG Ao

(Institute of Network Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China)

Abstract: A collaborative filtering recommendation algorithm based on news timeliness is proposed. Firstly, by analyzing the characteristics of the news timeliness, the timeliness model of news is established. Then, the user-based collaborative filtering algorithm is improved combining the news timeliness model. Finally, the experimental results show that this method can highly enhance the performance of user-based collaborative filtering algorithm, and ameliorate the accuracy and recall rate of news recommendation.

Key words: collaborative filtering; news recommendation; timeliness model

随着移动网络技术的高速发展, 网络新闻生产、传播速度都呈爆炸性的增长, 人们逐渐从新闻信息匮乏的时代进入了新闻信息过载的时代. 无论是新闻消费者还是新闻生产者都遇到了很大的挑战: 从消费者角度来看, 如何从大量新闻中发现自己感兴趣的新闻是一件非常困难的事情; 从生产者角度来看, 如何提高新闻浏览量和受众规模, 也是一件很困难的事情. 与此同时, 随着社会节奏的加快, 新闻消费者倾向于在更加碎片化的时间内浏览新闻, 因此新闻消费者主动搜索新闻以解决信息过载问题的意愿也就更低, 换言之, 用户希望在花费更少时间的前提下得到更适合自己的新闻.

个性化推荐系统[1]就是解决这一问题的重要工具. 在新闻推荐领域, 个性化新闻推荐系统通过联系用户和新闻, 一方面帮助用户发现对自己有价值的新闻, 另一方面让新闻能够展现在对它感兴趣的用户面前, 从而实现新闻消费者和新闻生产者的双赢. 其中协同过滤算法[2]是推荐系统中应用较为广泛的推荐算法, 该算法基于邻域; 根据领域选取的区别, 可以分为基于用户的协同过滤算法[3]以及基于物品的协同过滤算法[4]. 而在新闻推荐算法领域, 一般更适合使用基于用户的协同过滤推荐算法, 因为在一个新闻系统中, 用户量是相对固定且变化不明显的, 维护基于用户的协同过滤算法在性能上有更好的表现.

但基于用户的协同过滤算法容易忽视新闻信息的特性, 导致推荐新闻的时效性不足, 降低了新闻推荐的实际接受率; 并且需要不断调整相似用户的新闻信息

① 收稿时间: 2017-09-10; 修改时间: 2017-09-30; 采用时间: 2017-10-12; csa 在线出版时间: 2018-03-12

表, 在数据量大时, 算法时间开销会非常大.

本文针对上述问题, 提出了基于新闻时效性的协同过滤推荐算法, 该方法充分考虑到了新闻信息老化[5]的特点, 通过建立新闻的时效性模型, 改进了基于用户的协同过滤系统中对最近邻用户的选择; 在维护用户相似度的矩阵时, 将新闻集进行提前过滤, 保留时效性较高的新闻信息. 在新闻信息量较大的情况下也能维持算法的高性能, 本文利用该改进的算法, 对某网络新闻系统的新闻、用户行为数据集进行了仿真实验, 证明了本文所提方法的有效性.

1 新闻时效性模型和基于用户的协同过滤推荐算法

1.1 新闻时效性模型

对于推荐系统而言, 不同类型的物品具有不同的生命周期, 即它们的时效性会有很大的差别. 例如新闻信息就要比电影的生命周期短很多; 用户可能会满意对很久之前电影的推荐, 因为电影的信息熵并不会因为时间的推移而减少; 而对用户推荐老旧新闻, 很多时候都是无效的, 因为新闻的时效性非常重要; 即使是比较重要的历史性新闻信息, 实质上也算是过期信息, 对用户浏览新闻并无帮助.

从新闻信息的产生, 推荐, 成为热点, 衰退到最后的消失, 新闻信息在时间轴上总是呈现一定的规律; 这一点和应用信息计量学中的文献老化理论是相似的, 例如文献[6,7]就通过文献老化模型来描述网络信息的效用变化. 因此我们可以根据信息老化的特点, 建立新闻推荐系统的时效性模型, 定义如下:

定义 1. 新闻发布时刻t_n. 新闻发布时刻是指当新闻被生产完毕并正式发布的时间节点, 但还没被个性化推荐系统加入相似新闻集. 一般而言, 这个时刻与推荐系统更新时刻越近则说明新闻越新, 在经过初级过滤系统时, 它被过滤的可能性越低.

定义 2. 推荐算法更新时刻t_a. 由于新闻会不断发布, 因此推荐算法更新新闻信息库的时间也要根据系统的计算能力与实际新闻发布数量来设定, 一般而言, 两次更新间隔越短, 推荐效果越好, 但会消耗很大的系统性能, 需要根据实际系统来设定这个参数.

定义 3. 新闻生命周期T_s. 生命周期是指在新闻发布后, 自某个时刻起不再有新闻消费者对其做出消费行为, 从这个时刻 (t_n+T_s) 起, 零星的阅读行为可以认为是系统噪声不予考虑.

根据文献信息老化规律模型, 即贝尔纳在 1958 年提出的信息老化的负指数模型:

$$C(t)=ke^{-at} \tag{1}$$

公式 (1) 中, t为文献的出版年龄, $C(t)$表示t年时文献被引用频率. k是一个与文献分类有关的常数, a为文献的老化率.

将此模型应用于新闻信息, 可以用公式 (2) 定义, 将在后面验证负指数模型对新闻老化规律模型的适用性.

$$S(t_n,t)=e^{-a(t-t_n)+b} \tag{2}$$

其中, t表示当前时刻, t_n为新闻发布时刻, $S(t_n,t)$表示新闻在 t 时刻时的用户反馈统计数量, 需要根据新闻系统设定合适的时间粒度, 来统计该时刻分段的用户反馈数量. a是该新闻的老化系数, 表示该新闻随着时间推移其效果的衰减系数, a越小, 说明这条新闻的时效性越强; b则为某一个常数, 和新闻在发布后的初始统计时间粒度内的用户反馈统计数量有关, 以第一个统计时间段$t=t_n$为例, 在第一个时间粒度内, 用户反馈统计数量$S(t_n,t)=e^b$, b参数受新闻初始阅读量影响较大, 但不会表现新闻时效性的变化趋势; 即b较大的新闻会拥有比较大的初始阅读量, 但不能保证衰减速度慢; 在本文研究的时效性改进算法中, 更关心新闻本身的时效性变化趋势, 所以进行线性回归分析, 计算出a、b后, 主要采取a作为推荐算法的输入, 实际上对公式 (2) 变形后, 可以在公式 (3) 中更清楚的认识到老化参数与时间推移之间的关系:

$$S(t_n,t)=e^b e^{-a(t-t_n)} \tag{3}$$

变形后, 对老化曲线拟合更贴近于常见的负指数模型, 在数学表达上也更加直观. 只要求出老化系数a即可.

对于一条特定的新闻而言, 老化系数a代表了其在发布后传播效果的衰减速度, 发布后衰减不明显的新闻是时效性较强的新闻, 新闻消费者认为其有很高的时效价值; 而当一条新闻到达自己的生命周期T_s后, 那么此新闻就没有时效价值了.

1.2 基于用户的协同过滤算法原理

基于用户的协同过滤算法 (以下简称 UserCF 算法) 是推荐系统中最常见的算法之一, 应用十分广泛; 主要包括两个步骤, 首先寻找与目标用户兴趣度相似的用户集合. 然后找出在这个用户集合中的用户喜欢而目标用户尚未关注的物品信息, 将其推荐给目标用户.

在计算两个用户的兴趣相似度时, 主要利用行为的相似度来计算兴趣的相似度. 给定用户u和用户v, 集合$N(u)$表示用户u曾经有过正反馈的物品集合, 集合$N(v)$为用户v曾经有过正反馈的物品集合, 就可以计算u和v的兴趣相似度w_v^u,常见的相似度计算公式有两种, 第一种是杰卡德相似度计算公式[8]:

$$w_v^u=\frac{|N(u)\cap N(v)|}{|N(u)\cup N(v)|} \tag{4}$$

第二种是余弦相似度计算公式:

$$w_v^u=\frac{|N(u)\cap N(v)|}{\sqrt{N(u)||N(v)}} \tag{5}$$

根据用户之间的兴趣相似度, 可以给用户推荐和他兴趣最相似的K个用户喜欢的新闻, 公式 (6) 度量了 UserCF 算法中用户u对新闻n的感兴趣程度.

$$p(u,n)=\sum_{v\in S(u,K)\cap U(n)} w_{uv}r_{vn} \tag{6}$$

在此式中, $S(u,K)$包含了与用户u兴趣最为接近的K个用户, $U(n)$是对新闻 n 有过行为的用户集合, w_{uv}是用户u和用户v的兴趣相似度, r_{vn}代表用户v对新闻n的交互评分, 由其阅读时长、新闻长度、点赞、评论等正反馈行为归一化得出, 以衡量用户对新闻的兴趣度. 计算出$p(u,n)$后, 比较通用的做法是根据某个用户u对新闻n的预测评分$p(u,n)$, 做 Top-N 推荐, 即推荐出前N个高评分结果.

2 基于新闻时效性的协同过滤算法

2.1 算法设计

传统的 UserCF 算法中没有考虑信息时效性的问题, 这种做法可能适合电影、电商类系统, 但却忽略了新闻信息的时效性. 为了解决这个问题, 本文结合上文中的新闻时效性模型, 根据新闻衰老系数与生命周期, 对推荐新闻预测评分进行加权, 同时过滤过时新闻, 降低了算法输入数据的规模, 从而提高了推荐的效果.

2.2 数据模型

(1) 新闻信息集合, 表示新闻系统中的新闻集合, 会根据生产者的输出而更新, 加入新的新闻. 用集合$N=\{n_1,n_2,\cdots,n_i\}$来表示i条新闻的集合, 这个集合用来计算用户相似度.

(2) 推荐系统输入新闻集合, 表示新上架的新闻, 作为推荐系统在选取推荐新闻时的输入集; 同时也用来计算新闻时效性. 用集合$M=\left\{m_i|t_a<t_{n_{m_i}}+T_{s_{m_i}}\right\}$来表示. M集合包含了那些在推荐算法更新输入时刻仍然较为活跃, 并未消耗完生命周期的新闻, 这些新闻仍然有不少用户在阅读; 由于 UserCF 算法要进行比较复杂的用户偏好度计算, 同时要进行时效性模型的检验, 所以将输入集限定在仍然在生命周期内的新闻, 可以大幅降低算法复杂度.

(3) 用户信息集合, 表示新闻系统中的用户集合, 会根据系统用户的增加而更新, 但变化很缓慢. 用集合$U=\{u_1,u_2,\cdots,u_k\}$来表示k个用户的集合.

(4) 用户, 新闻兴趣矩阵R_{i*u}, 用户会对新闻做出不同的行为, 如阅读、点赞、分享及评论等; 通过用户行为偏好向量空间模型来计算用户对新闻兴趣值r_{un}.

(5) 新闻访问表, 用于记录新闻被阅读、收藏、赞操作及其具体时间, 用于计算新闻被访问随时间变化的趋势, 即其时效性.

2.3 通过时效性参数改进推荐算法

算法详细步骤如下:

1) 在推荐算法更新时刻t_a, 确定M集合, 并对M集合中的元素进行统计, 根据系统设定的时间粒度, 从新闻访问行为表中计算在每个时段内新闻被访问的数据. 通过时效性模型曲线拟合计算新闻老化率a, 并计算M集合中新闻老化率a的最大值, 记为$a_{\max}$.

2) 用户相似兴趣度计算, 可以选择余弦相似度计算, 也可以选择杰卡德相似度计算. 在选定相似度公式后, 计算用户相似兴趣度的输入集包含整个新闻集合N与用户集合U, 而非新闻子集合M, 这是为了获取用户完整的历史兴趣爱好, 以构建正确的用户相似兴趣度矩阵. 在这里我们选取杰卡德相似度公式:

$$w_v^u=\frac{|N(u)\cap N(v)|}{|N(u)\cup N(v)|} \tag{7}$$

3) 基于时效性参数改良预测评分结果:

$$p(u,n)=\sum_{v\in S(u,K)\cap U(n)} w_{uv}r_{vn}\left(1-\frac{a_n}{a_{\max}}\right) \tag{8}$$

对M集合中的新闻应用公式 (8), 计算出用户对M集合中新闻的兴趣度, 通过加入时效性参数作为加权, 即可得出最终预测兴趣评分.

4) 选取 Top-N 作为最后的推荐方法, 选择M集合中最高兴趣评分的N个元素作为用户推荐的结果.

3 实验分析

本文对某新闻报业集团的网络新闻数据与用户数

据进行实验. 共计 5436 条新闻, 43 187 个用户. 部分新闻数据如表 1 所示, 该表表示某个新闻的阅读情况.

表 1 实验部分新闻访问数据

采样时间	阅读量	阅读增量
2017-05-04 11:12	639	639
2017-05-04 11:42	1213	574
2017-05-04 12:12	1678	465
2017-05-04 12:42	2139	461
2017-05-04 13:12	2434	295
2017-05-04 13:42	2647	213
2017-05-04 14:12	2877	230
2017-05-04 14:42	2987	110
2017-05-04 15:12	3164	177
2017-05-04 15:42	3354	190
2017-05-04 16:12	3407	53

3.1 时效性模型实验

首先对新闻集进行时效性模型检验, 根据其阅读量变化, 使用负指数模型进行拟合校验, 以单条新闻为例, 图 1 即单条新闻阅读变化量拟合结果.

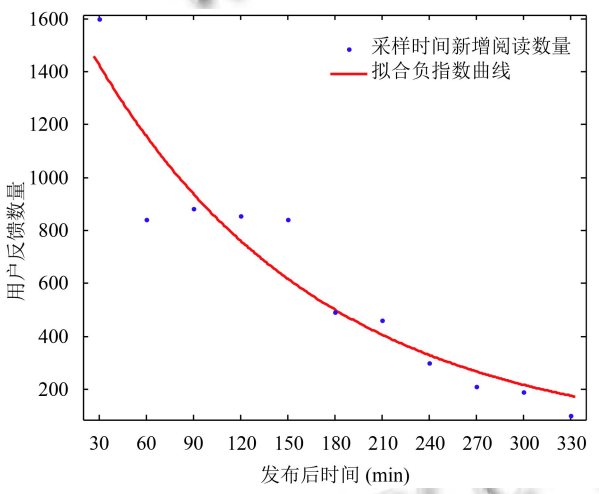

图 1 单条新闻拟合结果

表 2 为其时效性模型检验结果. 对于本文的时效性模型而言, SSE 与 RMSE 与模型输入数量级有关, 对结果没有太大的解释意义; R-square 与 Adjusted R-square 则表达了拟合结果与目标模型的效果, 越接近 1 说明模型越好. 表 2 中对应新闻浏览量变化趋势比较符合时效性模型曲线, 即阅读量与时间之间有较强的相关关系, 可以根据其拟合得到的老化系数作为推荐模型的输入.

表 2 某新闻拟合结果

SSE	R-square	Adjusted R-square	RMSE
2.055e+005	0.8945	0.8828	151.1

表 3 所示为对测试集合中的新闻进行时效性模型检验, 从中可以看出, 拟合结果 R-square 大于 0.80 的比例为 70.7%, 即其与负指数时效性模型拟合度较好, 有较高的说服度, 这些新闻的主要特点是平均总阅读量比较高, 因此噪音表现不明显; 而剩余的新闻平均总阅读量比较低, 受关注度低, 即使在发布的第一时间, 也很少有用户关注; 对在发布后的某个特定时间段的抗噪音能力较差, 用户反馈统计数量容易受波动, 难以体现时效性变化的总体趋势.

表 3 新闻集合拟合结果统计

R-square	新闻数量	占比(%)	平均总阅读量
≥0.80	3841	70.7	3944
<0.80	1595	29.3	906

针对时效性模型的拟合误差, 需要根据具体情况设定误差实验分析; 总体而言, 对于阅读量较大的新闻, 时效性模型是比较适用的, 拥有比较好的正确率与精度, 也可以从中看出新闻阅读量与时间的相关性; 对于抗噪声能力较差的非热门新闻, 可以将其剔除出新闻输入集合, 只对订阅用户推送, 因此就不会受到非精确时效衰减率影响.

3.2 基于时效性模型改进 UserCF 算法的实验

根据实验 (1) 的结果, 得出了时效性模型的新闻老化系数, 改进 UserCF 算法; 本实验采用 Mahout[9]作为 UserCF 框架, 在计算用户兴趣度时, 考虑时效性模型中老化系数的影响.

将实验 (1) 中的新闻数据集中 R-square 超过 0.80 的新闻集单独抽出, 形成一个新的新闻集 A, 与原新闻集 N 分别作为新闻总集, 进行两次实验, 实验步骤相同, 输入不同, 其他参数一致. 最后与传统的 UserCF 算法作为对照比较.

本实验中, 将用户行为数据集随机且均匀分成 8 份, 6 份作为训练集, 2 份作为测试集, 通过准确率与召回率来评价推荐算法的效果:

式 (9) 为准确率的公式:

$$Precision = \frac{\sum_u |R(u) \cap T(u)|}{\sum_u |R(u)|} \tag{9}$$

式 (10) 为召回率的公式:

$$Recall = \frac{\sum_u |R(u) \cap T(u)|}{\sum_u |T(u)|} \tag{10}$$

其中$R(u)$为推荐新闻集合, $T(u)$为用户实际喜欢的新闻信息. 准确率用于描述最终推荐列表中实际发生

的用户-新闻兴趣评分, 召回率用于描述有多少比例的用户-新闻兴趣评分出现在最后的最终的推荐列表中.

表 4 推荐算法实验结果

算法	总新闻集	准确率(%)	召回率(%)
UserCF	N	23.14	15.66
UserCF	A	29.32	18.31
时效性模型改进的UserCF	N	57.04	32.59
时效性模型改进的UserCF	A	66.40	30.14

表 4 为推荐算法实验结果, 首先对于不同的输入新闻集合 N 与 A, A 代表那些和时效性模型拟合程度较好的新闻数据集, 它们的平均阅读数量也较高, 抗噪声能力强; 实际上在一个推荐系统中, 热门新闻受推荐的概率会更大, 因此其准确率与召回率都会更高. 在传统的 UserCF 算法中, A 集合有着更高的推荐准确率与召回率, 但与 N 集合差别并不大, UserCF 算法没有考虑到时间推移的影响, 只响应了高阅读量新闻的特性, 略微提高了推荐准确率与召回率.

在相同的输入集合下, 改进后的 UserCF 算法有着更好的表现, 这是因为基于时效性模型改进的 UserCF 算法考虑到了新闻时效性衰减的因素, 较好地利用了老化系数, 降低了衰减较快新闻的权重; 对于和时效性模型拟合较好的新闻集合, 改进后的 UserCF 性能还会有提高, 这是因为这类新闻不仅衰减速度慢, 时效性强, 同时自身阅读量高, 故而推荐效果还会有所增强.

3.3 时效性模型误差分析

在时效性模型实验实验中, 针对拟合误差较大的新闻集合, 即 R-square 值低于 0.80 以下的新闻集合 B, 进行误差分析与算法性能分析.

表 5 误差实验结果

算法	总新闻集	准确率(%)	召回率(%)
UserCF	N	23.14	15.66
UserCF	B	19.28	13.31
时效性模型改进的UserCF	N	57.04	32.59
时效性模型改进的UserCF	B	34.40	19.68

表 5 为误差新闻集实验结果. 对于 R-square 值较低的新闻集合, 由于其和时效性模型偏差较大, 拟合所得的老化率a不能完全反应新闻实际失效的速度, 在引入 UserCF 算法时并不会大幅提高 UserCF 性能; 但从实际结果来看, 时效性模型改进的 UserCF 算法仍然有 15% 的提高. 这是因为高误差新闻集合中的新闻阅读量较少, 这些新闻很容易在算法的第一步时因为生命周期消耗殆尽, 从而被提前过滤. 没有进入输入集合中; 但相对于总新闻集合中的推荐实验结果, 这些误差使得时效性模型改进的 UserCF 算法的提高较为有限.

4 结论与展望

随着网络新闻量的爆发式增长, 如何在信息过载的情境下为新闻消费者提供合理的新闻推荐集成为了重点问题. 本文结合文献信息老化模型, 应用于新闻信息上, 利用时效性模型中的老化参数改进了基于用户的协同过滤算法, 对新闻-用户数据集进行了分析和研究. 从实验结果上看, 这种改进型算法适用于抗噪能力强的热门新闻, 能提高新闻推荐算法的准确率和召回率.

但是本文提出的方法只考虑了协同过滤算法中兴趣评分处理与新闻时效性的问题, 并没有解决协同过滤算法中冷启动[10]问题, 同时也没有考虑到系统初始状态下的用户稀疏性[11]问题. 所以该算法在未来还有很大的改进空间.

参考文献

1 Liu JX, Tang MD, Zheng ZB, *et al*. Location-aware and personalized collaborative filtering for web service recommendation. IEEE Transactions on Services Computing, 2016, 9(5): 686–699. [doi: 10.1109/TSC.2015.2433251]

2 马宏伟, 张光卫, 李鹏. 协同过滤推荐算法综述. 小型微型计算机系统, 2009, 30(7): 1282–1288.

3 项亮. 推荐系统实践. 北京: 人民邮电出版社, 2012: 33–37.

4 邓爱林, 朱扬勇, 施伯乐. 基于项目评分预测的协同过滤推荐算法. 软件学报, 2003, 14(9): 1621–1628.

5 陆研, 毛健骏, 屠方楠. 网络信息老化规律研究——新浪新闻与新浪微博实证研究. 高等函授学报 (哲学社会科学版), 2011, 24(12): 52–55. [doi: 10.3969/j.issn.1007-2187.2011.12.021]

6 马费成, 望俊成. 信息生命周期研究述评 (Ⅰ)——价值视角. 情报学报, 2010, 29(5): 939–947.

7 鞠菲. 网络信息老化实证研究——以新浪新闻为例. 情报杂志, 2010, 29(10): 41–45, 40. [doi: 10.3969/j.issn.1002-1965.2010.10.010]

8 李斌, 张博, 刘学军, 等. 基于 Jaccard 相似度和位置行为的协同过滤推荐算法. 计算机科学, 2016, 43(12): 200–205. [doi: 10.11896/j.issn.1002-137X.2016.12.036]

9 Owen S. MAHOUT 实战 (图灵程序设计丛书). 王斌, 韩冀中, 万吉译. 北京: 人民邮电出版社, 2014: 28–73.

10 孙小华. 协同过滤系统的稀疏性与冷启动问题研究[博士学位论文]. 杭州: 浙江大学, 2005.

11 林建辉, 严宣辉, 黄波. 融合信任用户的协同过滤推荐算法. 计算机系统应用, 2017, 26(6): 124–130. [doi: 10.15888/j.cnki.csa.005805]