DOI:10.16644/j.cnki.cn33-1094/tp.2016.12.006

图书推荐算法综述*

傅汉霖[1]，顾小宇[2]

(1. 东南大学成贤学院，江苏 南京 210088；2. 江苏省电力公司检修分公司)

摘　要：图书馆的信息库中保存着大量的读者检索信息和借阅记录，充分利用这些信息并结合高效的图书推荐算法可以充分地满足读者的借阅需求。综述了目前常用的图书推荐算法的思想、特点及应用，根据对图书馆的适用性分析了各推荐算法的优缺点，并提出了将适用性广泛的协同过滤算法推广为多特征的混合推荐算法策略的研究方向。

关键词：图书推荐算法；协同过滤；多特征；综述

中图分类号：TP391　　**文献标志码：**A　　**文章编号：**1006-8228(2016)12-21-03

A review of books recommendation algorithms

Fu Hanlin[1], Gu Xiaoyu[2]

(*1. Southeast University Chengxian College, Nanjing, Jiangsu 210088, China; 2. State Grid Jiangsu Electric Power Company*)

Abstract: There are a lot of readers' retrieval information and borrowing records in the library information database, to fully utilize the information and combine with the high efficient books recommendation algorithm can fully meet the needs of readers. This paper reviews the thoughts, characteristics and application of the current commonly used books recommendation algorithms, according to the applicability of the library, analyzes the advantages and disadvantages of each recommendation algorithm, and proposes a research direction to extend the collaborative filtering algorithm to the multi-feature mixed books recommendation algorithm.

Key words: books recommendation algorithm; collaborative filtering; multi-feature; summary

0 引言

在每座图书馆的数据库中都保存着大量的读者检索信息和借阅记录，这些信息蕴含着读者对馆藏图书资源的需求，使用相关推荐算法可以对数据库中所保存的读者检索信息和借阅记录进行归纳和整理，有助于图书馆进一步对馆藏资源的优化，预测读者对图书信息的潜在需求，也有助于图书馆个性化服务的智能化。本文参考了国内外关于图书推荐算法的相关文献，从多角度探讨了现有算法的核心思想，以推动图书馆相关推荐算法的深入研究。

1 推荐算法综述

1.1 基于密度的协同过滤算法

该算法的核心思想是：根据图书的归还时间，利用模糊理论的隶属函数来计算读者对图书的兴趣程度，并筛除读者不感兴趣的借阅记录。读者借阅的图书信息可以反映其偏好的图书类别；读者归还已借阅图书的时间可以反映读者对该书的兴趣；读者对续借的书一定是感兴趣的。

1.1.1 图书归还

图书归还时间可以反映读者对所借阅图书的偏好程度。如果刚借图书就立即归还，说明读者对该书不感兴趣；如果图书被续借，则表明对该图书感兴趣。

定义图书归还集：$R_{time}=\{r_1,r_2,\cdots,r_n\}$。

$$r_i=\frac{return_{time(i)}-borrow_{time(i)}}{T}$$

其中，$borrow_{time(i)}$、$return_{time(i)}$分别为读者对某册图书的借阅时间和归还时间，T为图书借阅规定还书周期。

收稿日期：2016-9-27

*基金项目：东南大学成贤学院2016年大学生实践创新训练计划(ycx1608)

作者简介：傅汉霖(1996-)，男，江苏南京人，本科生，主要研究方向：软件工程。

1.1.2 模糊值函数定义

用隶属函数u_{like}和$u_{dislike}$分别表示对图书感兴趣与不感兴趣的模糊程度，$f_{like}(r_i)$为读者基于图书归还的感兴趣与不感兴趣的模糊值。

$$f_{like}(r_i)=\begin{cases}1 & c<r_i<r_{max}\\ (ri-a)/(c-a) & a\le r_i\le r_{max}\\ 0 & r_{min}\le r_i<a\\ \perp & r_i<d_{min}\end{cases}$$

其中，a和c是隶属函数u_{like}和$u_{dislike}$的界定参数值。图书的归还时间区域r_a–r_{min}用来筛选出不感兴趣的书目，归还时间区域r_{max}–r_c筛选出感兴趣的书目。如果借阅时间超过规定的归还周期，则该图书的借阅时间信息无效。

1.1.3 推荐估值

读者借阅记录的聚类区域反映了读者对图书的兴趣特征，采用距离平方反比进行推荐：距离越近者，权重值越大。根据距离点q_0的最近质心所属的聚类区域，点q_0的推荐估值定义为

$$f_{like}(p_0)=\alpha_{CR}(i)\times\sum_{j=1}^{x}\beta_{CR(i),j}\times f_{like}(p_j)$$

其中，$\alpha_{CR}(i)$为点q_0最近聚类区域的区域权重；权重$\beta_{CR(i),j}=f(d_{0,j})/\sum_{j=1}^{m}f(d_{0,j}),\ f(d_{0,i})=1/d_{0,j}^2,\ d_{0,j}=\mathrm{distance}(q_0,P_{CR(i),j})$为点$q_0$聚类区域中离点$q_0$最近x个点的欧式距离[1]。

1.2 基于中图分类法的推荐算法

该算法的核心思想是：根据中图分类法和图书的特征向量计算图书的相似性，依据读者的特征向量和借阅记录计算读者的相似性，对其进行加权，产生最终推荐结果。

1.2.1 基于中图分类号的图书相似性

中国图书馆分类法，简称中图分类法，具有从总到分、从一般到具体的特点。采用汉语拼音字母与阿拉伯数字相结合的混合号码[2]。中图分类法把属于相同学科、具有相同主题的图书归为一个类。依次归类后，相似度最高的图书的分类号处于中途分类树的底层。

所以如果要比较两本图书的相似性，应先比较中图分类号最左边的字母。字母相同时，比较字母后的第一位数字，若相同，则比较第二位数字，以此类推。可构建目标读者r未外借图书i与已外借图书j基于中图分类号特征向量的相似度$sim_L(i,j)$：

$$sim_L(i,j)=\begin{cases}\dfrac{Layer(i,j)-1}{Layer(all)} & c(i)\ne c(j)\\ 1 & c(i)=c(j)\end{cases}$$

其中，Layer(i,j)为图书i的分类号c(i)与图书j的分类号c(j)在中图分类树中最近的父节点所在的层数；Layer(all)为中图分类树的总层数。

1.2.2 图书的受欢迎程度

设页数为Page(i)的图书j的所有读者集合为R(all)，读者r'∈R(all)，外借的日期为$Borrow_{date}(r',i)$和$Return_{date}(r',i)$，图书j基于页数特征向量的外借与归还时间间隔为：

$$Interval(r',i)=\frac{Return(r',i)-Borrow(r',i)}{Page(j)}$$

所有借阅过该图书的读者，平均每次从外借到归还图书的时间间隔为$A_{interval}(i)$。

设图书j入库的日期为Indate(i)，当前日期为Nowdate，Borrow(i)为该图书在时间段Nowdate–Indate(i)内被借阅的总次数。图书j在某时间段内平均被借阅次数ABorrow(i)。当前日期与出版日期之差的倒数衡量图书j的新旧程度New(i)，New(i)的值越大，表明图书i越新。

综上所述，可得出图书j受广大读者的欢迎程度Welcome(i)为：

$$Welcome(i)=\alpha\times A_{interval}(i)+\beta\times ABorrow(j)+\chi\times New(i)$$

其中，α、β、χ分别为该方程的系数[3]。

1.3 基于主题模型的推荐算法

该算法的核心思想是：通过对读者的历史借阅记录与其他图书数据进行相似度分析，得到与读者历史借阅图书相似度较高的图书；通过对读者的历史借阅记录与其他读者的历史借阅记录进行相似度分析，得到最近邻读者的历史借阅记录。通过求解图书被推荐的概率，最终得到读者潜在感兴趣的图书。

1.3.1 图书内容相似度

读者的历史借阅图书类别的集合$G=(g_1,g_2,\cdots,g_i,\cdots,g_l)$及每一类所对应的关键词集合$J=(j_1,j_2,\cdots,j_i,\cdots,j_l)$，其中$j_i=(m_1,m_2,\cdots,m_v)$。对于一本非目标读者借阅过的图书，可以根据图书对应的关键词集合与目标读者历史借阅记录中各类别的图书关键词进行相似度分析，得到：

$$sim_1=\left(\sum_{i=1}^{I}\sum_{r=1}^{v_i\cdot u}d_r\right)/\left(\sum_{i=1}^{I}\sum_{r=1}^{v_i\cdot u}d_0\right)$$

其中，v_i为目标读者历史借阅图书类别i的关键词个数。

由此可知，sim_i的值大则相似度越大，此图书被推荐的可能性也越大。取$d_0=1$，若$n_k=m_i$(即此图书的关键词与目标读者的历史借阅记录中某一类图书的关键词匹配)，则d_r取值为1，否则取值为0。

1.3.2 最近邻借阅者

通过其他借阅者的历史借阅记录，可能从中挖掘出目标借阅者新的感兴趣的图书。设有矩阵U(n,m)表示有n个目标读者与最近邻借阅者集合$P=(p_1,p_2,\cdots,p_n)$及m个图书集合$Q=(q_1,q_2,\cdots,q_n)$的评分矩阵，利用余弦相似度计算公式计算与读者相似程度较高的其他读者作为目标读者的最近邻。相似度计算公式如下：

$$sim_{\cos}(p_1,p_2)=\frac{\sum_{q\in Q_{1,2}}(U_{p_1,q}-\overline{U_{p_1}})(U_{p_2,q}-\overline{U_{p_2}})}{\sqrt{\sum_{q\in Q_1}(U_{p_1,q}-\overline{U_{p_1}})^2\sum_{q\in Q_2}(U_{p_2,q}-\overline{U_{p_2}})^2}}$$

其中，$Q_{1,2}$表示两个读者p_1,p_2具有共同评分的图书，Q_1为读者p_1有过评分的图书，Q_2为读者p_2有过评分的图书，$U_{p,q}$表示读者p_1对图书q的评分。$\overline{U_{p_1}}$与$\overline{U_{p_2}}$分别表示读者p_1,p_2对图书的平均评分。评分$U_{p,q}$的计算公式如下：

$$U_{p,q}=\frac{t_q-T_{\min}}{T_{\max}-T_{\min}}$$

其中，t_q为读者所花时间，T_{min}为统计开始时刻，T_{max}为统计结束时刻[4]。

2 结束语

基于内容相似度的推荐算法的推荐结果直观，但面对新用户和复杂情况无法对读者进行合适的图书推荐；协同过滤推荐算法的推荐个性化和自动化程度高[5]，但是面对新用户、新项目仍无法进行合适的图书推荐，对历史数据质量要求较高；基于中图分类法的推荐算法推荐结果直观，但个性化程度低。

将普通的协同过滤算法推广为多特征推荐算法，在此设计一种混合图书推荐策略，可以充分利用各种算法的优点，有关的研究表明这些混合算法的准确率要高于单独算法[6-7]。在该混合图书推荐策略中，若是新用户，则根据图书受大众读者欢迎程度对用户进行推荐，使得新用户即使刚使用系统，也可以获得推荐结果。用户开始借阅和检索图书，在数据库中留下历史借阅记录，可以基于中图分类法对用户进行相关书籍推荐。当用户的历史借阅记录达到一定数量时，可根据基于内容、最近邻读者、密度等算法进行有效的推荐。

参考文献(References):

[1] 武建伟，俞晓红，陈文清.基于密度的动态协同过滤图书推荐算法[J].计算机应用研究，2010.27(8).

[2] 国家图书馆《中国图书馆分类书》编辑委员会.中国图书馆分类法(5版)[M].北京图书馆出版社，2010.

[3] 李克潮，梁正友.基于多特征的个性化图书推荐算法[J].计算机工程，2012.38(11)

[4] 郑祥云，陈志刚，黄瑞，李博.基于主题模型的个性化图书推荐算法[J].计算机应用，2015.9.

[5] 陈永光.基于OPAC的高校图书馆个性化图书推荐算法研究[D].南京理工大学，2013.4.

[6] Soboroff I, Nicholas C. Combining content and collaboration in text filtering[C].ProcIn'l Joint Conf Artificial Intelligence Work-shop: Machine Learning for Information Filtering, Stockholm,1999:86-91

[7] Tran T, Cohen R. Hybrid recommender systems for electronic commerce. Proc. Knowledge-Based Electronic Markets[C].the AAAI Workshop. Menlo Park: AAAIPress, 2000:78-83