# 基于“读者—图书”二部图的个性化图书推荐方法

何绯娟,缪相林,许大炜,毕 鹏
(西安交通大学城市学院,陕西 西安 710018)

**摘 要**:个性化图书推荐已成为图书馆领域关注的热点问题,但面临着读者兴趣、图书内容难以获取以及“冷启动”等一系列挑战。文中基于图书借阅行为建立“读者—图书”二部图模型,并基于此提出个性化图书推荐方法。该方法首先根据书名计算图书之间相似度;其次,基于读者兴趣相似度对读者进行聚类,并生成每个读者的获选图书集合;最后计算每个读者与候选图书集合中每本图书的匹配度,并排序后输出推荐图书列表。实验结果表明,该方法能在未知读者兴趣、图书内容的情况下,有效地实现个性化图书推荐,并缓解了“冷启动”问题。

**关键词**:二部图;个性化推荐;图书;兴趣;冷启动
**中图分类号**:TP391.2 **文献标识码**:A **文章编号**:1673-629X(2015)05-0025-04
doi:10.3969/j.issn.1673-629X.2015.05.007

# Personalized Recommendation of Books Based on “Reader-Book” Bipartite Graph

HE Fei-juan, MIAO Xiang-lin, XU Da-wei, BI Peng
(Xi'an Jiaotong University City College, Xi'an 710018, China)

**Abstract**: Personalized book recommendations have become a hot area in library science. Current recommending methods, however, are facing the difficulty to automatically acquire reader interests and book topics, and the “cold start” problem. A novel personalized book recommending method based on “Reader-Book” bipartite graph derived from the book lending behavior is proposed. First, the semantic similarities among books are calculated utilizing the book titles. Second, readers are divided into different groups with the use of clustering analysis based on the similarity of reader interests. Every group is assigned a selected book set. Finally, each reader is recommended a preferable book list based on the matching degree between reader and book. Experimental results show that this method can recommend personalized books to a reader without knowing reader interests and book topics, and alleviate the “cold start” problem.

**Key words**: bipartite graph; personalized recommendation; book; interest; cold start

## 0 引 言

信息技术的不断进步极大地推动了图书馆服务向数字化、网络化以及普适化方向的发展,有效地解决了传统图书馆服务的时空延伸问题。然而,现有服务总体上仍采用“千人一面”的服务模式,不具有个性化、智能化的特点。以图书文献检索为例,现有服务通常被动接收用户提交的关键词或元数据,并通过字符串匹配机制向读者反馈结果列表;该过程未考虑读者自身的兴趣或偏好。这种服务模式不仅加重了读者在结果列表中选择所需图书的信息负载[1],而且也限制了图书文献的利用效率。

如何实现个性化图书推荐已成为图书馆领域关注的热点问题,同时也面临着诸多挑战,表现为:

(1)很难有效地获取读者的兴趣或偏好信息,且这类信息通常具有动态性(如学生的兴趣与当前所学课程紧密相关);

(2)通常只能获得图书的元数据信息,难以获取图书内容信息;

(3)个性化推荐中的普遍性难题,如数据稀疏、“冷启动”(cold-start)[2]等。

针对上述问题,文中建立了基于借阅行为的“读者—图书”二部图模型,提出了基于该模型的个性化图书推荐方法,并通过实证测试验证了所提方法的有

**收稿日期**:2014-06-30 **修回日期**:2014-09-30 **网络出版时间**:2015-04-22
**基金项目**:国家自然科学基金资助项目(61202184);陕西省教育专项项目(2013JK1207);陕西省教育科学规划课题(SGH13461);西安交大城市学院科研项目(2013KZ02,2014KZ02,2014KZ04)
**作者简介**:何绯娟(1977-),女,硕士,讲师,研究方向为数据挖掘。
**网络出版地址**:http://www.cnki.net/kcms/detail/61.1450.TP.20150422.1113.033.html

效性。

## 1 相关研究工作

目前,对于图书推荐的研究总体上可分为内容过滤(Content-based Filtering)与协作过滤(Collaborative Filtering)两类方法。

内容过滤是指利用推荐对象的特征与用户兴趣、偏好信息(或用户模型)的匹配度进行推荐[3]。典型研究如:Mooney 与 Roy 利用 Amazon 网站上的图书描述信息建立了 Naive Bayes 文本分类器用于向特定兴趣的读者推荐图书[4]。Tsuji 等结合图书借阅信息、书目信息提出了基于 Support Vector Machine(SVM)分类器的图书推荐方法[5]。基于内容过滤的方法通常受限于如何获取推荐对象特征与用户偏好信息。

协作过滤是指利用一组用户的偏好模式向特定用户进行推荐[6]。典型研究如:Sohail 等提出了基于观点挖掘(Opinion Mining)的图书推荐方法,该方法通过挖掘读者评价信息生成 Top10 的推荐列表[7]。Vaz 等提出一种混合协作过滤方法,该方法综合图书推荐与作者推荐的结果生成 Top-$n$ 的推荐列表[8]。目前,部分研究利用图书借阅日志建立二部图(Bipartite Graph)实现协作过滤[9-10],这些模型未考虑图书之间的相似性。协作过滤方法的优点在于不需要直接获取推荐对象的特征,但是存在"冷启动"问题,表现为部分图书(特别是新图书)缺少相关的借阅行为,很难对这部分图书进行推荐。

## 2 "读者—图书"二部图模型

图书馆信息系统中存储了关于图书借阅行为的大量历史数据。这些数据可以形式化表示为一个三元组集合。每个三元组的形式为 $(r,b,t)$,表示读者 $r$ 在 $t$ 时刻借阅了图书 $b$。三元组 $(r,b,t)$ 隐含了以下两方面含义:一是读者 $r$ 推荐图书 $b$;二是读者 $r$ 在 $t$ 时刻的兴趣是图书 $b$ 相关的主题。

由于读者兴趣的动态性,为此,需将三元组集合按时间戳 $t$ 划分为多个子集,每个子集对应一个时间窗口(设为 360 天),其中读者兴趣可当作是固定的。为了描述每个三元组子集,进一步建立了如图 1 所示的"读者—图书"二部图("Reader-Book" Bipartite Graph,RB-Graph)模型。

RB-Graph 模型可表示为一个二元组 $(R \cup B, E_{RB})$。其中,$R \cup B$ 为节点集,$R=\{r_i\}_n$ 与 $B=\{b_k\}_m$ 分别表示读者集合与图书集合;$E_{RB} \subseteq R \times B$ 为边集,表示特定时间窗口中涉及的借阅行为。对于读者 $r_i \in R$,设 $B_i=\{b_k \mid (r_i,b_k) \in E_{RB}\}$,表示 $r_i$ 借阅的图书集合。对于图书 $b_k \in B$,设 $R_k=\{r_i \mid (r_i,b_k) \in E_{RB}\}$,表示借阅 $b_k$ 的读者集合;若 $R_k=\varnothing$,表示图书 $b_k$ 在时间窗口内未被借出过。

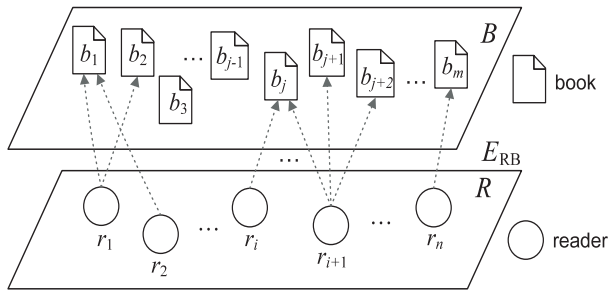


图 1 "读者—图书"二部图模型

基于 RB-Graph 模型,引出以下两个假设:

假设 1:对于图书 $b_k \in B$,若节点度(Degree)(即 $|R_k|$)越大,则 $b_k$ 越适合被推荐。

假设 2:对于读者 $r_i, r_j \in R$,若 $B_j$ 与 $B_i$ 越相似,则两者兴趣越接近,且 $r_j$ 借阅的图书越适合向 $r_i$ 推荐($B_j$ 与 $B_i$ 的相似度计算见 3.1 小节)。

以下基于 RB-Graph 模型与假设 1、2,提出一种个性化图书推荐方法。

## 3 个性化图书推荐方法

个性化图书推荐问题可表述为:对于读者 $r_i \in R$,生成 $B_i'(B_i' \subseteq B$ 且 $B_i' \cap B_i=\varnothing)$,满足对于 $\forall b_k' \in B_i'$,$d(r_i,b_k') \geqslant \max\limits_{b_k \in B-B_i-B_i'} d(r_i,b_k)$,其中,$d(r,b)$ 是读者 $r$ 与图书 $b$ 的匹配度函数。

基于 RB-Graph 模型,提出了"图书相似度计算—兴趣社区发现—匹配度计算"三阶段的个性化图书推荐方法。在图书相似度计算阶段,根据书名计算图书之间相似度。在兴趣社区发现阶段,计算读者所借图书集合的相似度,以此作为读者兴趣相似度对读者进行聚类,生成每个读者的获选图书集合。在匹配度计算阶段,计算每个读者与其候选图书集合中每本图书的匹配度,并排序后输出推荐图书列表。

### 3.1 图书相似度计算

图书内容信息通常情况下很难获取,为此,利用图书名称计算图书之间相似度。通过对 2 872 本计算机类图书的分析发现,图书名称平均仅有 7.8 个字,是一种典型的短文本。由于存在特征稀疏问题[11-12],传统的 Bag-of-Words 与 $n$-gram 模型都难以用于表示图书名称。这里,采用文献[13]提出的方法计算图书 $b_x, b_y$ 的相似度 $s(b_x,b_y)$。

$$s(b_x,b_y)=\frac{|W_x \cap W_y|}{|W_x \cup W_y|} \tag{1}$$

式中,$W_x$ 与 $W_y$ 分别是利用图书 $b_x, b_y$ 书名中的关键词在 Wikipedia 中搜索出的页面标题集合。显然,

$s(b_x,b_y)$ 越大,图书 $b_x,b_y$ 的相似度越高。

### 3.2　兴趣社区发现

兴趣社区发现中,需要计算读者之间的兴趣相似度。根据假设 2,对于读者 $r_i,r_j \in R$,其兴趣相似度 $S(r_i,r_j)$ 定义为所借图书相似度的均值。

$$S(r_i,r_j)=\frac{\sum_{b_x\in B_i,b_y\in B_j} s(b_x,b_y)}{|B_i||B_j|} \tag{2}$$

利用式(2)定义的读者兴趣相似度,采用 FGKM 增量式聚类算法[14]对读者进行聚类,该算法能满足借阅行为与用户兴趣的动态性需求。聚类后生成 $R$ 的一个划分 $R=\{R_l'\}_p$。对于 $r_i \in R$,若 $r_i \in R_l'$,则其候选的推荐图书集合 $B_i'$ 定义如式(3),即同簇内其他读者借阅的图书集合与本人借阅的图书集合的差集。

$$B_i'=\bigcup_{r_k\in R_l'} B_h - B_i \tag{3}$$

### 3.3　“读者—图书”匹配度计算

对于读者 $r_i \in R_l'$,对候选推荐图书集 $B_i'$ 中的每本图书计算匹配度,并按匹配度输出 Top-$k$ 项的推荐结果。

根据 RB-Graph 模型的两个假设,对于图书 $b_k' \in B_i'$,读者 $r_i \in R$ 与之的匹配度 $d(r_i,b_k')$ 定义为 $R_j'$ 内其他借阅 $b_k'$ 的读者与读者 $r_i$ 的相似度之和。

$$d(r_i,b_k')=\frac{\sum_{r_g\in R_l\wedge r_g\neq r_i\wedge(r_g,b_k)\in E_{RB}} S(r_i,r_g)}{|R_l'|} \tag{4}$$

式中,$(r_g,b_k')\in E_{RB}$ 表示 $r_g \in R_l'$ 借阅过图书 $b_k'$;$|R_l'|$ 用于对匹配度进行归一化。显然,对于 $b_k'$,若簇 $R_l'$ 内借阅该书的读者越多,$d(r_i,b_k')$ 越大,符合假设 1;若簇 $R_l'$ 内借阅 $b_k'$ 的读者与 $r_i$ 越相似,则 $d(r_i,b_k')$ 越大,符合假设 2。

根据 $d(r_i,b_k')$ 对 $B_i'$ 中的图书进行排序,则 Top-$k$ 的图书集合为 $B_{i,k}' \subseteq B_i'$。

设 $B_0=\{b_i \mid R_i=\varnothing \wedge b_i \in B\}$ 表示时间窗口内未被借阅的图书(通常为新书集合),为了解决针对 $B_0$ 中图书推荐的“冷启动”问题,将 $B_0$ 中与 $B_{i,k}'$ 最相似的图书也作为推荐图书,即推荐的图书 $b_0 \in B_{i,k}'$ 为:

$$b_0=\arg\max_{b_x\in B_0}(\max_{b_y\in B_{i,k}'} s(b_x,b_y)) \tag{5}$$

由式(5),最后推荐图书的集合为 $B_{i,k}' \cup \{b_0\}$。

## 4　测试与验证

利用四川省某大学图书馆的图书数据(涉及 26 309种图书)与该馆 2011 年间的图书借阅日志数据(http://www.datatang.com/data/45757,涉及 59 111 条借阅日志)进行实验。借阅日志中,只利用“学号”(代表读者)、“借书日期”、“题名”(即书名)三个字段。基于上述日志数据,建立了借阅行为的二部图,该图包含 26 309 个图书节点、24 500 个读者节点以及 59 111 条表示借阅行为的边,其中 11 706 个图书涉及借阅行为,其余图书节点为孤立节点(可看作未被借阅过的图书)。

表 1　推荐图书评估

| 学号 | 借阅图书 | 推荐图书 | $P@5$ |
|---|---|---|---|
| 084309xxxx | 现代韩国语;新概念韩国语 | 新韩国语基础教程;韩国语实用语法;新编初级韩国语;初级韩国语;韩国语入门 | 1.0 |
| 094300xxxx | 机械设计;聚酰亚胺新型材料;分离膜制备与应用 | 机械设计学习指南;聚酰亚胺:化学、结构与性能的关系及材料;新型分离技术;机械设计习题与解析;机械设计与机械原理考研指南 | 0.6 |
| 104102xxxx | 货币银行学 | 现代货币、银行与金融市场;货币金融学;商业银行经营学;中央银行学;货币金融学 | 0.8 |
| 078502xxxx | 内科学试题库;CT 诊断与临床 | 新编内科学应试向导;实用内科学;内科学复习多选题;X 线 CT 诊断学图谱;实用 CT 诊断图谱 | 0.6 |
| 084101xxxx | 中国美术史 | 中国美术史;中国美术史教程;中国美术史图像手册;中国美术全集;中国美术 | 1.0 |
| 084101xxxx | 图解园林植物造景;住宅建筑 | 园林设计初步;园林景观快题设计;建筑画环境表现与技法;20 世纪住宅建筑;都市住宅 | 0.8 |
| S08xxxx | 数据挖掘导论;决战恶意代码;Head First 设计模式 | 数据挖掘:概念与技术;软件剖析:代码攻防之道;C#设计模式;数据挖掘原理;机器学习 | 0.8 |
| S09xxxx | 空间的语言 | 建筑空间论:如何品评建筑;Java 程序设计语言;建筑方案设计;城市空间美学;十字街头的语言文字 | 0.6 |
| S11xxxx | 图像处理 Photoshop 7.0 入门与提高 | Photoshop CS2 从入门到精通;ANSYS 11.0 基础与典型范例;Adobe Photoshop CS 中文版经典教程;Photoshop CS 中文版照片处理应用 100 例;中文版 Photoshop CS 标准培训教程 | 1.0 |
| S11xxxx | 资本论 | 《资本论》劳动价值论的具体化;货币与资本市场;《资本论》历史典据注释;国际金融学概论;马克思的历史、社会和国家学说 | 0.8 |

由于无法通过问卷调查方式判断推荐图书的有效性,故以人工方式判断推荐图书的有效性,并采用 $P@5$ 指标对推荐效果进行量化评估[15]; $P@5$ 定义为前 5 个推荐图书中与推荐对象兴趣相关的图书所占比重。选择 10 个读者进行测试,结果如表 1 所示(学号后四位做了隐私处理)。该结果验证了所提方法的有效性,并支持对未被借阅过图书的推荐。

## 5 结束语

文中基于图书借阅行为建立了“读者—图书”二部图模型,基于该模型提出了一种个性化图书推荐方法。实证测试表明,该方法能在未知读者兴趣与图书内容的情况下,有效地实现个性化图书推荐,并缓解了协作过滤类推荐方法中普遍存在的“冷启动”问题。下一步研究主要包括:一、“读者—图书”二部图是一个随着图书借阅行为动态变化的网络,需要研究如何基于该网络实现“读者—图书”匹配度的并行化增量式计算;二、研究图书推荐中的相关度、多样性等多指标排序算法。

**参考文献:**


[1] 曾庆辉,邱玉辉. 一种基于协作过滤的电子图书推荐系统[J]. 计算机科学,2005,32(6):147-150.

[2] 王显飞,陈 梅,李小天. 基于约束的旅游推荐系统的研究与设计[J]. 计算机技术与发展,2012,22(2):141-145.

[3] 刘平峰,杨 柳,朱孔真. 一种基于内容过滤的服务资源推荐新技术[J]. 武汉理工大学学报:信息与管理工程版,2014,36(5):668-672.

[4] Mooney R J,Roy L. Content-based book recommending using learning for text categorization[C]//Proceedings of the fifth ACM conference on digital libraries. San Antonio, Texas, USA:ACM,2000:195-204.

[5] Tsuji K, Takizawa N, Sato S, et al. Book recommendation based on library loan records and bibliographic information [J]. Social and Behavioral Sciences,2013,147:478-486.

[6] 纪良浩. 协作过滤信息推荐技术研究[J]. 重庆邮电大学学报:自然科学版,2012,24(1):78-82.

[7] Sohail S S,Siddiqui J,Ali R. Book recommendation system using opinion mining technique[C]//Proc of the international conference on advances in computing,communications and informatics. Mysore,India:[s. n. ],2013:1609-1614.

[8] Vaz P C,de Matos D M,Marings S,et al. Improving a hybrid literary book recommendation system through author ranking [C]//Proceedings of the 12th ACM/IEEE-CS joint conference on digital libraries. Washington,DC,USA:IEEE,2012:387-388.

[9] 李树青,徐 侠,许敏佳. 基于读者借阅二分网络的图书可推荐质量测度方法及其个性化图书推荐服务[J]. 中国图书馆学报,2013,39(3):83-95.

[10] Wang F S,Yang H Y. Books-borrowing behavior in library management system[J]. Complex Systems and Complexity Science,2012,9(1):55-58.

[11] Chen M,Jin X,Shen D. Short text classification improved by learning multi-granularity topics[C]//Proceedings of the twenty-second international joint conference on artificial intelligence. Barcelona,Catalonia,Spain:AAAI Press,2011:1776-1781.

[12] 邱云飞,王琳颖,邵良杉,等. 基于微博短文本的用户兴趣建模方法[J]. 计算机工程,2014,40(2):275-279.

[13] Banerjee S,Ramanathan K,Gupta A. Clustering short texts using Wikipedia[C]//Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval. Amsterdam,Netherland:ACM,2007:787-788.

[14] Bagirov A M,Ugon J,Webb D. Fast modified global k-means algorithm for incremental cluster construction[J]. Pattern Recognition,2011,44(4):866-876.

[15] 薛源海,俞晓明,刘 悦,等. 信息检索中的带权邻近度度量研究[J]. 计算机研究与发展,2014,51(10):2216-2224.


(上接第 24 页)


[6] Kotonya G,Sommerville I. Requirements engineering:processes and techniques[M]. [s. l. ]:John Wiley & Sons,1998.

[7] Pressman R S. Software engineering:a practitioner's approach [M]. 7th ed. [s. l. ]:McGraw-Hill Companies Inc,2010.

[8] 张国生. 基于层次着色 Petri 网的需求工程过程框架[J]. 计算机应用与软件,2011,28(8):17-19.

[9] Institute of Electrical and Electronics Engineers. IEEE standard glossary of software engineering terminology (IEEE Std 610.12-1900)[S]. New York:IEEE,1990.

[10] Gotel O,Finkelstein A. An analysis of the requirements traceability problem[C]//Proceedings of the first IEEE international conference on requirements engineering. Los Alamitos: IEEE Computer Society Press,1994:94-101.

[11] Wieringa R. An introduction to requirements traceability[R]. [s. l. ]:Vrije University,1995.

[12] Pohl K. Requirements engineering:fundamentals, principles, and techniques[M]. Berlin:Springer-Verlag,2010:610-613.

[13] Domges R,Pohl K. Adapting traceability environments to project specific needs[J]. Communications of ACM,1998,41(12):54-62.

[14] Ichikawa A,Yokoyama K,Kurogi S. Control of event_driven systems-reachability and control of conflict-free petri nets [J]. Trans Soc Instrum Control Eng,1995,21(4):324-330.

# 基于“读者-图书”二部图的个性化图书推荐方法



作者： 何绯娟， 缪相林， 许大炜， 毕鹏， HE Fei-juan， MIAO Xiang-lin， XU Da-wei， BI Peng

作者单位： 西安交通大学城市学院,陕西 西安,710018

刊名： 计算机技术与发展 ISTIC

英文刊名： Computer Technology and Development

年，卷(期)： 2015(5)