

基于用户日志的OPAC推荐系统 设计与实现*

唐乐 李向前

(西南交通大学图书馆, 成都 611756)

摘要: 为提高OPAC检索的效率和图书资源的利用率, 本文以西南交通大学图书馆的OPAC日志数据为基础, 通过对该数据的采集、处理, 以及采用基于统计特征的算法对数据分析和推荐过程进行建模, 完成OPAC推荐, 并且给出具体的系统设计、实现过程和应用效果。用户在OPAC的检索效率和图书借阅率都有提高, 系统功能满足设计要求。

关键词: 日志分析; OPAC; 推荐

中图分类号: G251

DOI: 10.3772/j.issn.1673-2286.2019.01.004

OPAC是查询图书馆馆藏资源的门户, 学界对OPAC检索相关问题的研究非常多, 主要分为两个方面。①对检索方式的研究。如姜婷婷等^[1]对OPAC系统的用户搜索行为进行了详细的统计分析, 吴丹等^[2]对用户检索式构造行为及相邻检索式重构行为进行了深入分析, 鲁丹^[3]的OPAC搜索日志可视化系统采用了Mysql存储日志数据、JS图表框架进行数据的可视化展示。②对检索结果相关内容推荐的研究。如黎邦群^[4]介绍了一种基于检索行为图书推荐的相关图书推荐方法。通过对相关文献查阅, 笔者发现通过日志分析来研究用户检索行为的文献较多, 但基于此对OPAC检索相关推荐的研究较少, 同时缺少关于系统如何完成自动化的、可视化的日志管理工作以减少人工操作方面的研究。本文以西南交通大学图书馆(以下简称“本馆”)的OPAC系统为例, 对用户行为日志数据进行较为深入的分析, 并且给出构建推荐模型和系统实现的方法。

1 需求分析

本馆的OPAC系统由馆员自主研发, 共享图书馆自

动化系统的数据与相关功能, 采用联盟馆模式, 在系统中可以检索到图书馆和各个院系资料室的馆藏资源。该OPAC系统有分馆、资料类型、检索项、关键字等4个检索项, 其中前3项条件默认为“全部内容”, 即默认条件下用户只需要输入关键字就可以进行检索。检索结果记录为单位显示简要的书目信息, 用户可以按书名、作者、出版社、书目号、索书号及ISBN号进行二次检索, 也可以根据书名和作者字段对检索结果进行排序和筛选。同时, 用户可以点击记录标题来查看书目的详细信息和复本馆藏信息, 并进行后续的预约或预借操作。

现有的OPAC功能比较传统, 如果用户要快速找到需要的图书, 必须对检索方法及图书信息非常了解, 才能通过构造比较准确的检索式来实现快速检索。实际上, 大部分用户既不能够熟练地使用检索系统也对目标图书不太了解, 他们必须尝试不同的检索式来缩小检索范围, 以找到一个相对匹配的结果集。而结果集中的记录也通常并非按用户的偏好排序, 用户还要不断地翻阅检索记录来找到所需的图书。这样的OPAC显然无法满足用户简单快速查询信息的需求。同时, 大量有价值的图书资源因未能被发现而长时间闲置, 这也给图书

*本研究得到四川省文化厅图书情报学与文献学规划项目“基于OPAC日志分析的用户检索行为研究”(编号: WHTTXSM[2016]37)资助。

馆造成很大的浪费。

为了使用户能够简单快速地找到自己感兴趣的图书，OPAC需要辅助用户的检索过程，智能地推荐相关信息，提升用户检索的效率和体验。本文讨论的推荐方式包括检索式推荐和书目推荐两种，即在用户检索操作时进行检索式推荐，提供相关关键字等推荐信息，辅助用户选择最佳的检索策略；在用户完成检索式并查询到检索结果后进行书目推荐，提供相关的热门图书，辅助用户选择图书。

2 研究方法

本研究的基础是OPAC的日志数据。用户在使用OPAC进行各种操作时会产生大量的OPAC日志数据，如果将这些数据收集、整理，并按一定的方法进行分析和挖掘，就能了解该用户的偏好，可以据此向用户推荐相关信息，进而提高检索的效率和质量。因此，本文对2018年1—10月本馆的OPAC日志数据进行收集、整理，并设计基于统计特征的推荐算法，对主要推荐过程进行建模，最终设计了一个自动化推荐系统来完成这些工作。

2.1 数据采集和处理

OPAC的日志数据包括检索行为数据和后续行为数据两种。检索行为数据包括时间、客户端信息、资料类型、检索项和关键字等，采集的重点是检索式的信息。检索后续行为有翻页、查看书目详情、借阅/预约信息等。后续行为数据是返回检索结果后，读者进行的查看书目详情或借阅等后续操作的相关记录。

虽然日志数据在采集时经过一定的筛选，但由于网络、浏览器和系统性能等原因，采集的数据可能存在不完整或重复采集等问题，这些无效的数据会影响数据分析的结果，造成推荐内容匹配度低。因此，在数据分析前，需要进行细致的数据处理。数据处理包括三方面的内容：一是删除不完整的数据；二是规范数据格式；三是甄别重复的用户请求，去掉由于误操作或网络爬虫产生的一些批量请求，进而保证数据的真实性。数据处理的内容见表1。

2.2 推荐过程建模

推荐过程建模是OPAC推荐系统开发的核心工作，

表1 数据处理的内容

无效数据	处理方法
内容不完整	删除相关的日志记录
格式不规范	修改数据为统一格式
重复的用户请求	屏蔽非校内客户端发出的请求日志； 检查日志记录的间隔时间，如果同一客户端的同一种操作（检索、翻页、点击书目等操作）小于3秒则认为是重复操作导致的，并删除重复的记录

直接关系到推荐内容的质量。主要流程包括识别对象的特征、通过推荐算法分析数据、将推荐结果呈现给用户。其中采用基于统计特征的推荐算法，所依赖的样本数据完全来源于用户的行为日志，当样本数据量足够大时，其统计特征会趋于稳定，由此建立推荐内容与用户检索行为间的关系。

推荐算法分为数据统计和内容推荐两个部分。无论是检索式推荐或是书目推荐，都要预先分析日志数据并保存统计结果，那么在用户检索时只需调用处理过的统计结果来生成推荐内容即可。这样将计算量大、耗时长的日志分析任务提前利用空闲时间处理，计算量小、实时性要求高的检索推荐任务进行实时处理，有助于提高OPAC推荐功能的实时响应速率和提升用户检索体验（见图1）。

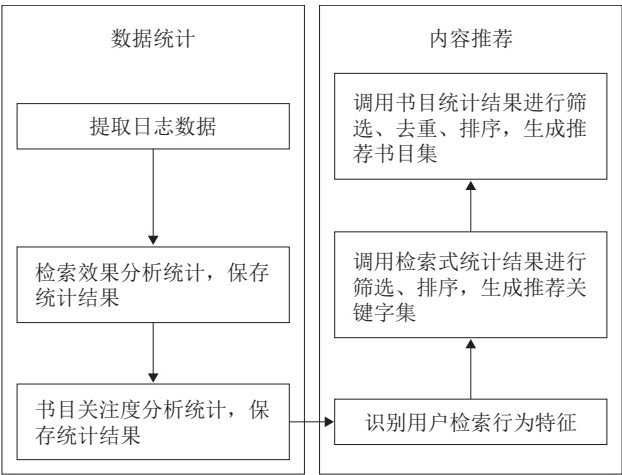


图1 推荐算法流程

2.2.1 数据统计

数据统计的第一步是提取日志数据，包括检索记录和书目查看记录。用户的每一次检索操作都会产生1条检索记录，记录中主要有分馆、资料类型、检索项、

关键字4个检索项。用户的每一次查看书目操作都会产生1条书目查看记录,记录中主要有用户使用的检索式和书目号。

第二步是对检索效果进行量化评分和统计。根据

用户检索满意度和获取结果的便捷度,制定检索效果评分标准(见表2)。评分标准涵盖了检索式的构造、书目详细记录浏览和图书借阅等多个方面。所有具体评分标准项可由用户自行设置,使其更加合理和贴近实情。

表2 检索效果评分标准

评分指标	指标说明	评分说明
是否返回了结果	检索请求是否有结果返回	没有返回结果总分评为0分,有返回结果评5分,最高5分
检索条件数量	用户检索时使用的检索条件数量	除关键字以外的检索条件1个扣3分,最低评0分,最高15分
用户查看书目情况	用户在第几页查看了哪些书目	第1页查看1种书目评5分,第2页查看1种书目评3分,第3页查看1种书目评2分,第4页至第10页查看1种书目评1分,10页以后评0分,最高35分
查看的书目借阅情况	用户查看的书目的借阅或预约情况	1种书目查看后1天之内借阅或预约加5分,最高45分
总分	-	以上指标分值累加,满分100分

检索效果分析流程如图2所示。首先按照检索效果评分标准,通过日志记录对每次检索操作的效果评分,计算方法见公式(1)。

$$score(Q) = \sum_{i=1}^n score(Z_i) \quad (1)$$

Q 表示一次检索记录, Z 表示1~ n 个检索效果的评分指标。然后判断已存在的统计样本库中是否存在该检索式的统计结果,不存在则增加新的检索式统计样本,随后将该检索式的分数汇总,计算该检索式的检

索次数和平均分数,通过加权计算出该检索式的综合评分,计算方法见公式(2)。

$$score(Q) = \frac{\sum_{i=1}^n score(Q_i)}{n} \times w_1 + count(Q) \times w_2 \quad (2)$$

Q 表示一种检索式, Q_i 表示该检索式的第 i 次检索记录, w_1 为平均分数的权重, w_2 为检索次数的权重, w_1 和 w_2 可由用户自行设置。最后,生成包含检索式及对应的检索次数、平均分数和综合分数的统计结果。

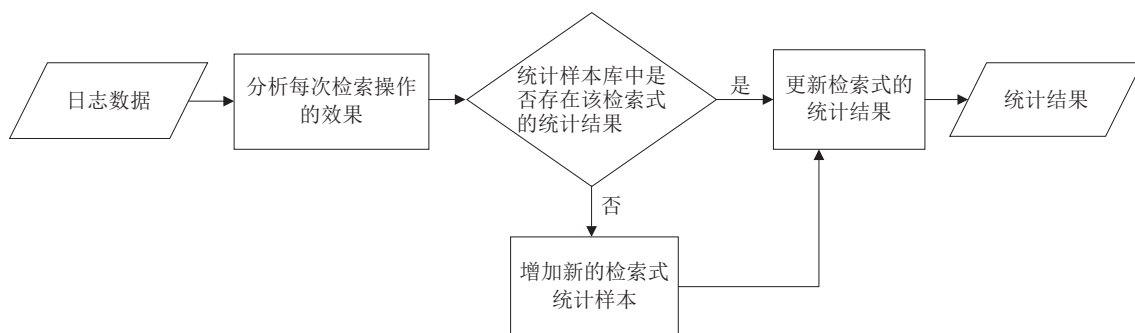


图2 检索效果分析流程

第三步是对书目的关注度进行量化评分和统计。根据用户后续的检索行为,如书目详细信息浏览情况和借阅情况,制定书目关注度的评分标准(见表3)。用户同样可自行设置具体评分标准项。

书目关注度分析流程是按照书目关注度的评价标准,通过日志记录分析每种书目的关注度。计算方法见公式(3)。

$$score(B) = \sum_{i=1}^n score(Z_i) \quad (3)$$

B 表示一次书目查看记录, Z 表示1~ n 个书目关注度的评分指标。然后判断已存在的统计样本库中是否存在该书目的统计结果,不存在则增加新的书目统计样本,随后将该书目的分数汇总,计算出查看次数和平均分数,通过加权计算出该书目的综合分数,计算方法见公式(4)。

$$score(B) = \frac{\sum_{i=1}^n score(B_i)}{n} \times w_1 + count(B) \times w_2 \quad (4)$$

B 表示一种书目, B_i 表示该书目的第 i 次查看记录, w_1 为平均分数的权重, w_2 为查看次数的权重, w_1 和 w_2 可由用户自行设置。最后, 生成包含用户使用的检索式、书目及对应的查看次数、平均分数和综合分数的统计结果。

表3 书目关注度评分标准

评分指标	指标说明	评分说明
用户查看书目情况	用户是否通过点击等操作查看书目及馆藏等相关信息	用户查看了书目评1分, 否则评0分, 满分1分
用户借阅书目情况	用户是否在查看书目后1天内进行了借阅或预约等有借阅意愿的操作	用户借阅或预约了图书评2分, 否则评0分, 满分2分
总分	-	以上指标分值累加, 满分3分

2.2.2 内容推荐

内容推荐的第一步是在用户进行检索条件的选择或填写时, 将这些操作信息发送至系统后台, 由系统后台进行识别。

第二步是检索式的实时推荐。根据用户的行为特征, 与统计结果进行比对。比对的方法是统计结果中检索式的分馆、资料类型、检索项3个检索条件与用户输入的相等, 并且统计结果中检索式的关键字等于或包含

用户输入的关键字, 则该检索式被筛选为备选内容。然后对筛选出的检索式按综合评分从高到低排序, 取排位靠前的检索式的关键字作为推荐内容。

如表4所示, 根据用户输入的检索式, 经过与统计结果比对, 筛选出3个备选检索式, 它们的分馆、资料类型、检索项3个检索条件与用户的检索式相同, 关键字都包含“C++”的内容, 按综合评分排序后生成的推荐关键字依次为“C++”“精通C++”“C++入门”, 用户可以参考推荐的内容, 选择满意的关键字进行检索。

表4 检索式推荐案例

	分 馆	资料类型	检索项	关键字	综合评分	排 序
用户检索式	西南交通大学图书馆	所有图书	书名	C++	-	-
备选检索式1	西南交通大学图书馆	所有图书	书名	C++	80	1
备选检索式2	西南交通大学图书馆	所有图书	书名	C++入门	50	3
备选检索式3	西南交通大学图书馆	所有图书	书名	精通C++	70	2

第三步是书目的实时推荐。在用户完成检索后, 根据用户使用的检索式, 与统计结果进行比对。比对的方法是统计结果中检索式的分馆、资料类型、检索项、关键字4个检索条件与用户的相等, 则该检索式对应的所有书目被筛选为备选内容。然后通过书名、ISBN等信息进行查重并删除重复的书目。最后, 对剩下的书目按综合评分从高到低排序, 取排位靠前的书目作为推荐

内容。

如表5所示, 根据用户使用的检索式, 经过与统计结果比对, 筛选出该检索式产生的检索结果中有3种查看过的书目, 按综合评分排序后生成的推荐书目依次为“C++ Primer Plus (第五版) 中文版”“C++ PRIMER 中文版: 特别版”“C++程序设计”, 用户参考推荐的书目进行后续的操作。

表5 书目推荐案例

检索式	书目号	检索项	ISBN	综合评分	排 序
分馆=西南交通大学图书馆 资料类型=所有图书 检索项=书名 关键字=C++	123	C++ Primer Plus (第五版) 中文版	978-7-115-13416-2	20	1
	345	C++程序设计	978-7-302-08599-4	13	3
	232	C++ PRIMER中文版: 特别版	978-7-115-22017-2	17	2

3 系统设计与实现

按照以上的研究方法,本研究设计了一个OPAC日志管理系统来完成上述工作。该系统包括日志采集器、日志分析工具、推荐内容接口和WEB管理界面4个部分。

用户在OPAC进行检索及后续操作时,会触发日志采集器采集用户的行为数据,并将数据提交至后台服

务器,存储在OPAC日志数据库中;日志分析工具定期从日志数据库和图书馆自动化系统数据库抽取相关数据进行数据处理和数据分析,处理记录和分析结果保存在日志数据库中;推荐内容接口负责识别用户的检索行为,实时调用统计结果进一步处理,并向OPAC返回推荐内容;WEB管理界面提供可视化的WEB界面,用户通过浏览器可随时查看OPAC日志分析系统的工作状态。系统的总体设计见图3。

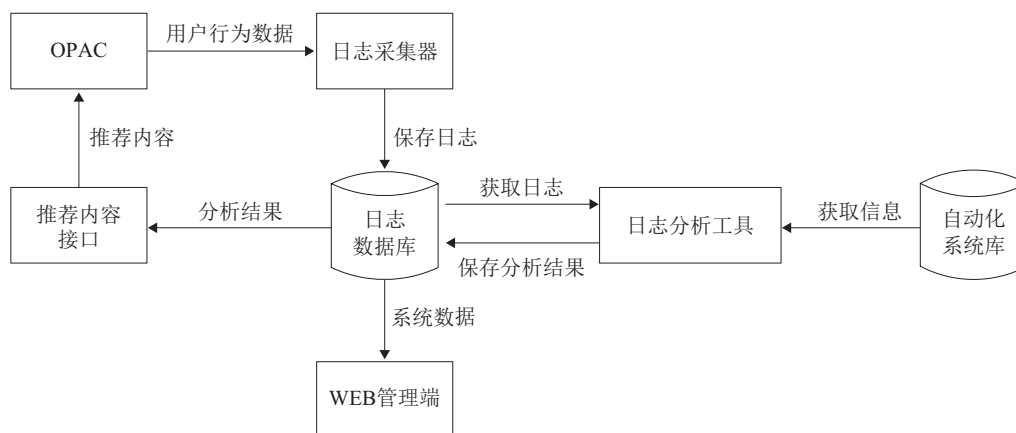


图3 系统总体设计

3.1 日志数据结构

系统存储的日志数据包括采集的OPAC用户行为数据和日志分析结果数据。用户行为数据包括检索日志表(见表6)和书目浏览日志表(见表7)。检索日志表主要存放用户的检索条件、返回的结果数量,每次检索会生成唯一的检索id,与书目浏览日志关联;书目浏览日志表主要存放书目所在的页码、书目号与对应的检索id。

表6 检索日志表

字段名称	字段类型	说明
time	datetime	访问时间
ip	varchar (30)	用户ip地址
branch	varchar (30)	分馆
zllx	varchar (20)	资料类型
category	varchar (20)	检索项
keyword	varchar (100)	关键字
record_count	integer	检索结果记录数量
id	varchar (20)	检索id

表7 书目浏览日志表

字段名称	字段类型	说明
time	datetime	访问时间
search_id	varchar (20)	检索id
brn	integer	馆藏书目号
page	integer	页码

日志分析后的结果分别记录在检索效果统计表和书目关注度统计表。检索效果表存放检索式的信息,以及该检索式的总评分和检索次数,每个检索式有唯一id标识,与书目关注度关联;书目关注度表存放书目号和检索式id,以及该检索式的检索结果中书目的关注度总评分和查看次数。

3.2 详细设计

3.2.1 日志采集器

很多OPAC日志分析的数据来源是应用服务器产生的日志数据,这种方式产生的日志数据无效请求很多,需要进行大量的数据清洗。而本文的日志数据通过

在OPAC页面嵌入脚本的方式进行采集,避免了许多无效请求,并且基本不影响现有的功能。日志数据实时地存入数据库,便于用户实时监控系统的工作状态。

3.2.2 日志分析工具

日志分析工具按事先制定的执行计划,定期执行相关任务,无须人工干预。该工具长期驻留在服务器中,提供用户界面,用户可在服务器中查看实时状况,也可在WEB界面中配置和查看执行历史和执行计划。日志分析工具包括数据分析和数据分析两个功能。

数据处理功能定期地扫描数据库中采集的原始数据,发现未经处理的数据,自动进行验证并做相应处理,处理完成后留下的合法数据做上“已处理”的标记,不合法的数据存入错误数据表中备查,每次数据处理的执行情况都保存到数据库中。用户可在WEB界面中查询错误的数 据,监控数据处理的执行情况,通过分析发现一些系统的问题并做相应的修改。

数据分析功能定期扫描数据库中经过处理的日志

数据,首先检查检索请求日志的内容,发现未分析的记录,自动获取该检索请求的后续行为数据,这些数据有翻页数据、书目浏览数据及借阅/预约数据。随后,按评价指标进行评分和结果汇总,结果数据保存至数据库中,并做“已处理”标记。

3.2.3 WEB管理界面

WEB管理界面包括运行监控、数据采集管理、数据处理管理、数据分析管理和系统设置等功能模块(见图4)。运行监控是当前系统运行的基本情况,让用户快速地监测系统故障和及时处理;数据采集管理模块可以查询系统采集的OPAC日志,并提供基本的统计功能;数据处理管理模块可以查询系统自动清洗掉的OPAC日志数据,并提供基本的统计功能;数据分析管理模块可以查询中间结果和最终结果数据,通过模拟的OPAC检索界面测试推荐功能和详细的推荐过程,并提供基本的统计功能;在系统设置模块中可以设置评价指标分数等参数,以及查询系统日志。

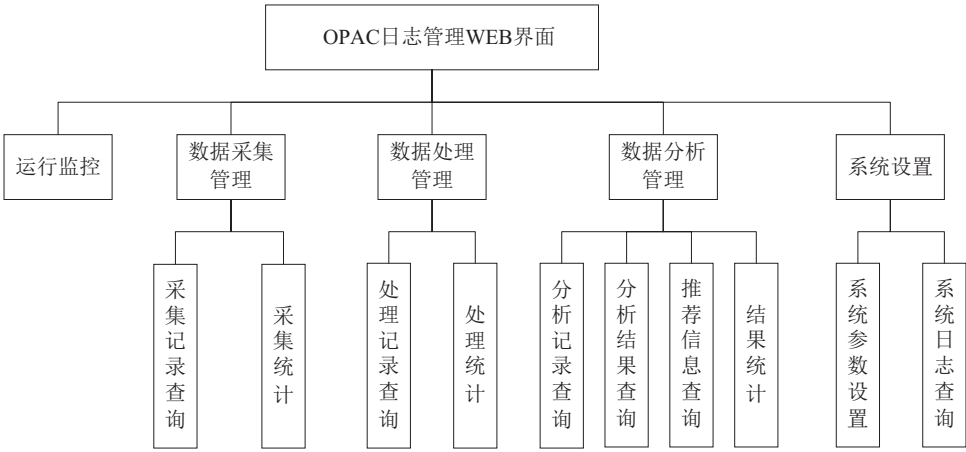


图4 WEB管理界面功能模块

3.2.4 推荐内容接口

OPAC的实时推荐功能通过推荐内容接口来完成,该接口接收来自OPAC的用户实时操作数据,从日志数据库中调用统计结果,并生成推荐内容返回到OPAC,包括检索式推荐接口和书目推荐接口两个部分。将系统实时推荐的功能设计成接口形式,可以方便OPAC在任何需要的地方接入推荐功能,提高系统设计和部署的灵活性。用户在检索时,会触发OPAC调用推荐内容接

口,并将推荐内容返回给用户。

3.3 系统实现及应用效果

系统部署在Linux服务器上,采用Mysql数据库存储数据,Java语言编写后台程序,并运行在Tomcat服务容器上。其中,日志分析工具采用Swing框架实现,具有可视化的界面;日志采集器前端采用Ajax跨域访问技术采集用户行为数据并发送至后台服务器,后台

为WEB接口形式,接收OPAC客户端发送的日志数据;WEB管理界面采用Ajax技术异步加载数据、ECharts框架生成图表等方法,实现不同数据的整合显示,提升用户体验;OPAC检索界面上,采用Ajax的跨域访问技术获取推荐数据,并通过页面下拉框和表格的方式呈现给用户。系统具有良好的可移植性,采用成熟的编程语言与开发平台,保证了日志采集和分析的稳定和高效。

2018年1—10月,系统采集到29万条检索数据、48万条翻页数据、39万条书目浏览数据,通过对日志数据的统计分析,生成11万条检索效果统计数据、13万条书目关注度统计数据。在OPAC推荐功能上线后,本文分别采集了上线前一周和上线后一周的OPAC使用数据,统计结果见表8。从基本的统计指标来看,检索请求略有下降,翻页次数降低约16.54%,表明用户不需要过多的翻页就能找到需要的图书,检索效率有一定的提高;图书借阅(预约)率提高22.14%,表明在OPAC中增加推荐内容后,用户更容易查到需要的图书,提高了图书利用率。从推荐的统计指标来看,用户推荐的检索请求占有检索请求的14.83%,查看推荐书目占有查看书目的17.85%,说明用户对推荐的信息有一定的需求。因此,统计结果表明本文的研究方法和系统设计符合要求,具有一定的应用价值。

4 总结与展望

针对本馆OPAC目前检索效率低、图书使用率低的问题,研究了基于用户日志对OPAC检索进行相关信息推荐的方法,设计并实现了一套完整的系统,验证了该方法的有效性。但由于数据量有限、研究时间短等原因,本研究对OPAC检索推荐的研究还不够深入,仍存

表8 应用效果统计结果

统计项	使用前	使用后	变化率
检索请求数量	9 912	9 112	-8.07%
用户翻页次数	17 855	14 901	-16.54%
浏览书目的数量	15 378	15 324	-0.35%
借阅(预约)率	11.88%	14.51%	22.14%
使用推荐的检索请求数量和百分比		1 351 (14.83%)	
查看推荐书目的数量和百分比		2 736 (17.85%)	

在一些不足待改进。①本文研究的对象主要是OPAC的日志数据,数据来源单一,其中参与分析的数据也较少,产生的分析结果可能存在误差或无法分析,造成推荐内容的选取和排序不准确。②研究方法的局限性。本文主要研究检索方式与检索效果间的关系,研究检索效果而对检索行为关联性的研究较少,需要进一步完善。下一步将扩大分析数据的来源,改进数据分析方法,提高数据挖掘的深度和广度,提高推荐内容的质量和OPAC的检索效率。

参考文献

- [1] 姜婷婷,王淼,高慧琴. OPAC系统用户搜索行为日志分析——以武汉大学图书馆为例[J]. 图书情报知识, 2015(5): 46-56.
- [2] 吴丹,董晶. 移动图书馆用户查询式构造行为分析[J]. 图书情报工作, 2016, 60(18): 7-13.
- [3] 鲁丹. OPAC搜索日志可视化系统的设计与实现[J]. 现代计算机(专业版), 2016(12): 67-70, 75.
- [4] 黎邦群. 基于检索行为的非个性化图书推荐[J]. 图书馆杂志, 2013, 32(8): 36-41.

作者简介

唐乐,男,1983年生,硕士,馆员,研究方向:数字图书馆,E-mail: tl2938@qq.com。
李向前,男,1985年生,硕士,助理馆员,研究方向:数字图书馆。

Design and Implementation of OPAC Recommendation System Based on Users' Log Analysis

TANG Le LI XiangQian
(Southwest Jiaotong University Library, Chengdu 611756, China)

Abstract: In order to improve the efficiency of Search and the usage of book resource, based on the log data of OPAC in Southwest Jiaotong University Library, completed OPAC recommendation by data collection, data processing and model the data analysis and recommendation process with the algorithm of statistical character, and give the specific system design, implementation process and application effect. User's search efficiency and book lending rate in OPAC were improved, and system functions meet design requirements.

Keywords: Log Analysis; OPAC; Recommendation

(收稿日期: 2018-12-26)