计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2021,30(8):111−117 [doi: 10.15888/j.cnki.csa.008096]

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

基于 Web 知识发现的图书数字资源个性化检索系统①

黄小根

(中共佛山市委党校, 佛山 528300)
通讯作者: 黄小根, E-mail: xxglhxg@163.com

摘　要: 在针对用户在 Web 上难以从海量的图书数字资源中找到符合需求的资料, 本文设计实现了基于 Web 知识发现的图书数字资源个性化检索系统. 该系统利用 Web 知识发现、智能代理、数据挖掘等技术, 设计出用户登录模型、用户兴趣生成模块、优化搜索结果等模块, 通过各模块的设计达成用户行为对兴趣度的影响, 个性化模型的更新, 以及搜索结果的处理, 进一步提升了 Web 上图书数字资源的检索质量, 期待通过本次研究, 为同领域内的图书数字资源个性化检索服务的构建, 提供一些有价值的参考资料.

关键词: Web 知识发现; 图书数字资源; 个性化; 检索系统

引用格式: 黄小根.基于 Web 知识发现的图书数字资源个性化检索系统.计算机系统应用,2021,30(8):111–117. http://www.c-s-a.org.cn/1003-3254/8096.html

Personalized Retrieval System of Digital Book Resources Based on Web Knowledge Discovery

HUANG Xiao-Gen

(Party School of the Foshan Municipal Committee of CPC, Foshan 528300, China)

Abstract: In response to users' difficulty in searching for what they need from massive digital book resources on the Web, this study develops and implements a personalized retrieval system of digital book resources based on Web knowledge discovery. The system uses Web knowledge discovery, intelligent agent technology, data mining, and so on to design modules such as user login model, user interest generation module, and optimized search results. This achieves the influence of user behavior on interest, the update of the personalized model, and the processing of search results, further improving the retrieval quality of digital book resources on the Web. We hope this research will provide some valuable reference for the construction of personalized retrieval services of digital book resources in the same field.

Key words: Web knowledge discovery; book digital resources; personalization; retrieval system

随着科学技术水平的不断进步, 网络上的图书数字资源越来越为丰富, 然而, 面对互联网上海量的图书数字资源, 依靠传统互联网的搜索引擎功能, 已经不能满足用户快速获取所需的知识. 由于现有搜索工具大部分没有考虑到用户的兴趣、个性特征及历史偏好, 往往使得用户的需求无法与意向的图书数字资源有较准确匹配, 也由此造成用户所搜集到的图书信息存在较大差异[1]. WWW (World Wide Web) 即全球广域网, 自身具有重复性、数量庞大, 以及无序性的特点, 已经成为当前图书数字资源检索的主要工具. 但随着用户的检索需求增加, Web 所反馈至用户的结果也越来越多, 在大量图书信息中, 用户也越来越难定位自身感兴

① 基金项目: 中共佛山市委党校《党校图书馆咨政信息服务平台》(GDJAFS2018031C)
Foundation item: "Party School Library Advisory and Political Information Service Platform" of Party School of Foshan Municipal Committee of CPC (GDJAFS2018031C)
收稿时间: 2020-12-11; 修改时间: 2021-01-11; 采用时间: 2021-02-02; csa 在线出版时间: 2021-07-31

趣的图书[2]. 因此, 打造基于 Web 知识发现图书数字资源个性化检索系统的设计, 准确获得搜索图书数字资源, 对于提高图书数字资源搜索性能, 有着重要的影响意义, 也是本次研究的重点.

1 基于 Web 知识发现图书数字资源个性化检索服务需求及关键技术分析

1.1 个性化检索的定义及服务需求

传统图书数字资源大多数是以结构化的格式存储在图书馆管理系统数据库中, 并以二维表结构表达图书信息, 但随着信息技术的发展, 图书数字资源表达形式变得越来越丰富, 大量电子书籍及资料由结构化向半结构化或非结构 (文档、图片、HTML、各类报表及音视频) 的格式存储在 Web 网络上, 由于非结构化的数据结构是不规则或不完整, 没有预定义的数据模型, 普通用户无法使用常规的检索方法找到关联性的知识. 因此, 如何在网络中从海量的、嘈杂的图书数字资源中找到满足特定的需求, 已经成为普通用户面临的一个难题, 而引入个性化检索服务是有效解决上述问题的有效方法.

个性化检索是基于 Web 知识发现、智能代理及数据挖掘等信息技术, 根据用户的特征、偏好、浏览记录及需求, 针对不同的用户采用不同的方式和策略, 提供不同的检索结果的综合技术服务[3]. 个性化检索服务是以用户需求为中心, 改变了以往无论哪个用户在检索平台以不同关键词搜索, 却得到相同结果的现象, 满足不同用户的不同需求, 提供了千人千面的个性化服务. 这种服务是在当用户不明确真实需求时, 个性化检索平台主动对用户的过往的历史检索记录、检索习惯和个体信息 (用户兴趣、性格、行为) 进行分析, 推测用户真实意图, 并利用 Web 知识发现、智能代理技术及数据挖掘等技术, 挖掘出隐含的、可利用的、有效的知识与用户需求进行关联匹配, 在完成对匹配结果进行索引、过滤及排序后, 主动地向用户推送其感兴趣或所需的图书数字资源信息[4].

个性化检索服务, 首先应该是要以用户需求为核心, 从信息输入到检索结果的呈现, 要以用户需求为主导, 提供友好、交互式的人机接口界面. 其次, 个性化检索服务要是一种个性化的服务, 在用户需求不明确的情况下, 善于挖掘用户有关信息, 精准推测出用户真实的需求, 从而主动地提供特定信息源. 最后, 个性化检索服务是一种智能化的服务, 充分利用 Web 知识发现、大数据、人工智能、云计算等信息技术, 结合用户的个体信息和偏好, 提供用户所需的信息资源和服务, 并根据信息资源的变化, 积极引导用户做出最佳选择.

1.2 关键技术分析

1.2.1 Web 知识发现

随着网络精准搜索要求越来越高, 过去由人工分类的方法已经无法满足现在的搜索分类需求. 在大部分网页无法达成精准搜索分类的情况下, Web 的自动分类功能是一种有效的解决途径[5]. 设计检索系统时, 可利用 Web 知识发现技术功能, 对图书数字资源所在的网页进行分类, 通过标引达成对图书数字资源所在网页的分类实施, 使标引与检索形成一体化, 并且该种检索还具备分类浏览的功能, 通过检索关键词, 直接标引的方式, 快速让用户获得所需要的图书数字信息[6].

由于图书数字资源主要汇集在 Internet 服务系统中, 通过对用户相关信息的搜集, 从而建立用户访问模式[7], 而 Web 知识发现主要是通过该工具的挖掘功能, 实现对用户感兴趣图书数字资源的快速获取建模的关键技术[8,9]. Web 知识发现是将 Web 本身具的挖掘功能, 应用于 Internet 个性化服务中, 利用 Web 知识发现技术, 达成 Internet 个性化服务, 更好的满足用户图书信息搜集的个性化需求, 其根本原理在于对图书资源信息的挖掘, 第一, 是针对网页内容进行分析, 采用 Web 知识发现的自动分类技术, 通过搜索功能, 进入图书数资源领域, 进行全面性的解析[10]; 第二, 是利 Web 知识发现针对用户访问过程中, 留下的日志进行数据挖掘, 从而对用户提供具有个性化的图书数资源检索[11], 第三, 是利用 Web 知识发现针对结构挖掘, 根据结构获取的图书信息, 由导航指引进入, 在此过程中对于一些图书资源库进行网站的设计, 并设有评价界面, 了解用户的满意程度, 通过图书主题或关键词的搜索, 打造个性的检索效果[12].

1.2.2 智能代理技术

在 20 世纪 90 年代, 智能代理的理论和技术已经被提出来, 它是一个涉及到人工智能、数据库技术及自然语言处理的计算机科学领域, 尤其在人工智能领域有较深度的应用. 智能代理技术具有智能性、主动性及适应学习性等特点, 一般使用智能代理技术处理复杂的数据分类、数据分析及数据信息加工.

智能代理技术在信息检索领域中利用智能代理服

务器收集用户需求, 根据用户定义的规则, 利用特有的通信技术协议向特定的用户推送信息. 在个性化检索的研究成果中, 充分体现智能代理技术得到广泛应用. 在一些没有特定要求的用户检索需求下, 可以将复杂的工作代替用户完成, 如图书数字资源信息的主题筛选、查询、管理等, 可通过智能代理技术推算用户可能产生的意图, 形成自主化的图书信息制定, 以及相应的资源调整, 并且制定可能需要的计划[13]. 在图书数字资源的个性化检索中, 通过智能代理技术, 可以有效的达成推理, 该技术自身的知识源非常丰富, 能够进一步的推测用户意图, 并将一些海量及复杂的图书信息快速整理, 按照用户的需求给予提供相应的接收, 设置相应的自动拒绝功能. 在该技术作用下, 也可以训练个性化检索模型, 提升检索功能, 进一步增强图书数字资源检索系统中处理问题的能力.

1.2.3 数据挖掘技术

数据挖掘技术是 Web 知识发现技术的一个分类, 一般认为数据挖掘是指利用决策树、神经网络、回归、关联规则、聚类等多方面的技术, 从海量的数据中抽取隐含的、未知的、可利用的信息, 并用于决策或知识存储的数据分析方法[14]. 数据挖掘一般用于事物描述和预测, 由于其具有聚类、数据关联分析及数据分类等特点, 经常被用于 Internet 上的图书数字资源的数据清洗、集成、变换、模型评估及知识表示. 另外, 数据挖掘技术可以被用于传统图书馆管理系统数据库的检索查询调用, 也可以用于非结构化的图书数字资源进行统计、分析及推理. 同时, 还可以利用检索图书数字资源信息挖掘事件之间的关联性, 从而对信息进行预测.

2 基于 Web 知识发现图书数字资源个性化检索系统的设计

2.1 系统总体设计

基于 Web 知识发现图书数字资源个性化检索系统设计, 可以有效的根据收集到用户信息, 用户在网页上的操作, 不断积累经验, 从而推测用户的兴趣和行为, 最终用户完成检索后, 快速反馈用户的需求结果. 基于 Web 知识发现图书数字资源个性化检索系统, 主要是增强了学习与更新的用户模型, 相较于以往系统的单纯检索更具智能化, 并且在 Web 知识发现的基础上, 进一步优化了查询与优化结果模块. 个性化检索系统的设计, 还进一步的考虑用户之间的差异, 利用 Web 知识发现提升图书数字资源个性化检索质量. 具体系统设计可见结构图 1 所示.

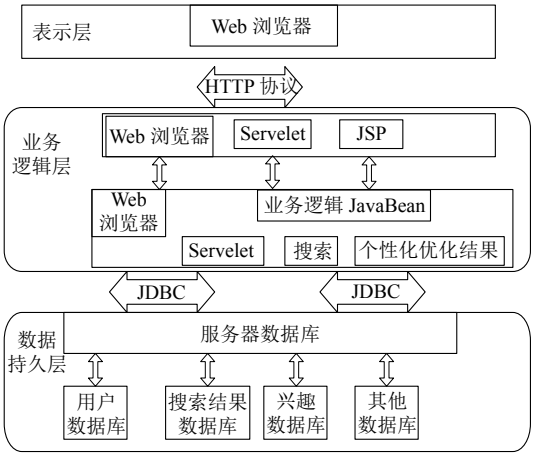

图 1 图书数字资源个性化检索系统结构图

由图 1 可见, 基于 Web 知识发现图书数字资源个性化检索系统结构图, 表示层为 Web 知识发现浏览器. 然后是业务逻辑层, 由应用服务器与业务逻辑共同支撑用户管理搜索模块、个性化模块、优化结果模块. 在数据持久层设置了用户数据库、搜索结果数据库、兴趣数据库、其他数据库, 所有数据库时时反馈相应信息, 并进行自我学习, 增强自身的知识, 实时反馈业务逻辑层相关数据处理信息, 达成图书数字资源个性化检索系统的设计效果.

2.2 图书数字资源个性化检索系统的设计流程

用户数据登录流程具体见图 2 所示.

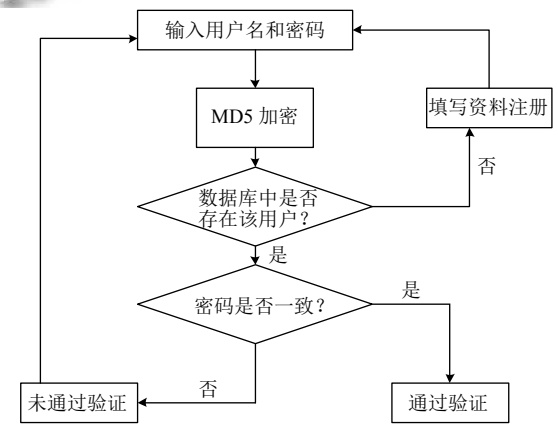

图 2 用户数据登录流程图

由图 2 可见, 系统会接收到来自 Web 浏览器端的用户搜索请求, 根据搜索请求调用相关业务, 对数据持

久层的图书数字资源进行有效的访问, 并且对所需要处理的数据返回至浏览器端, 随即快速反馈用户的请求. 这种个性化的系统设计主要是依据用户在图书数字资源库浏览网页页面中的满意度所反馈的信息, 了解用户的兴趣 (见图 3 所示), 再进一步的优化搜索结果, 反馈用户感兴趣的网页图书数字资源内容. 用户在登录网页过程中需要完成注册, 也可以访客的身份进入浏览, 而已注册的用户身份, 可以完成搜索功能, 在注册与登录过程中, 均设置了 MD5 的加密技术, 从而保证用户信息的安全性.

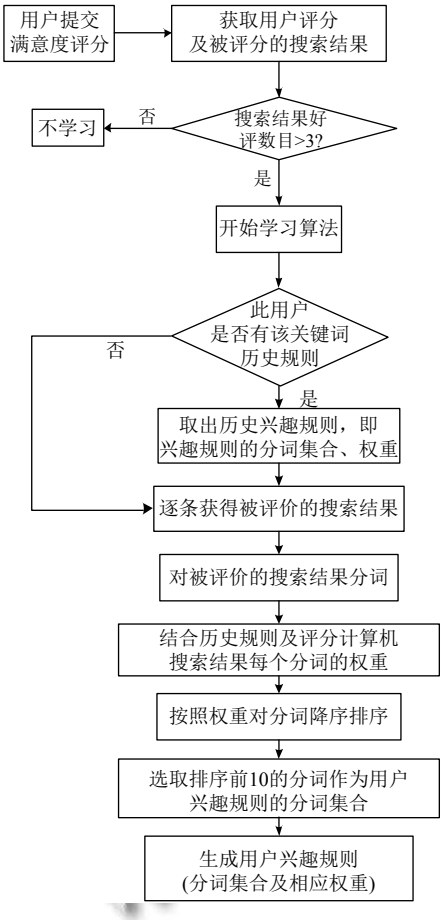

图 3 用户兴趣生成流程图

由图 3 可见, 用户在图书数字资源的浏览页面可以提交满意度评分, 系统根据提交的满意度评分获取用户的搜索结果, 根据用户满意度评分, 了解用户对图书数字资源的兴趣, 再进一步的为用户优化搜索结果. 具体可见图 4.

2.3 系统数据库设计

图书数字资源个性化检索系统的一切功能实现, 均需要获取用户的信息, 只有在了解用户信息的基础之上, 才能够对相关信息进行综合处理分析, 达成个性化的服务需求. 由此在系统数据库设计方面, 主要设计用户需求数据、搜索图书数字资源表、用户兴趣表、用户与规则关联表、搜索结果分值表.

用户需求数据表设计见表 1 所示.

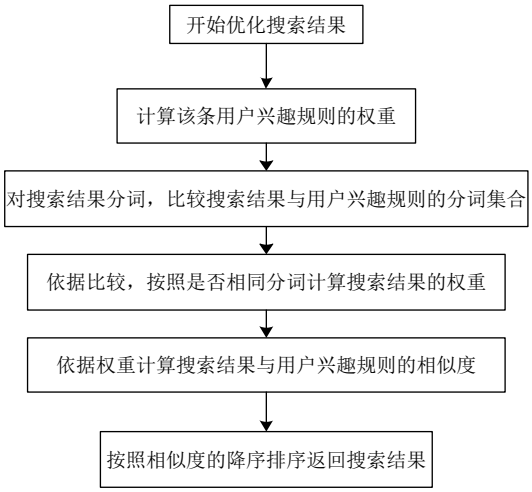

图 4 优化搜索结果流程图

表 1 用户表

字段名	是否为主键	字段类型	字段大小	说明
User name	是	Int	11	用户名称
User Password	否	Varchar	20	用户密码
User mailbox	否	Varchar	20	用户邮箱
Usermobile number	否	Varchar	15	用户手机号
Registration Time	否	Varchar	15	注册时间
Remarks	否	Varchar	15	备注

搜索图书数字资源表主要来保存图书具体类别和类别详细信息, 具体数据表设计见表 2 所示.

表 2 搜索图书数字资源表

字段名	是否为主键	字段类型	字段大小	说明
Search Number	是	Int	11	搜索编号
Book Resources Category	否	Varchar	10	图书资源类别
Digital Library Resources	否	Varchar	10	图书数字资源
Key words	是	Varchar	10	关键词
Number of results	是	Varchar	10	结果数
Time	是	Varchar	10	时间

用户兴趣规则表主要是通过用户满意度评分, 进一步获得用户对于图书数字资源信息的兴趣. 具体设计内容见表 3 所示.

用户与规则关联表主要是通过用户与获得的兴趣

关联, 进一步获得用户对于图书数字资源的个性化需求, 更好的提出优化结果. 具体设计内容见表 4 与表 5 所示.

表 3 用户兴趣规则表

字段名	是否为主键	字段类型	字段大小	说明
Rule number	是	Int	11	规则编号
Keyword rules	否	Varchar	20	关键词规则
Page number generated by rule	否	Varchar	10	规则产生的页码
When the rule was generated	是	Varchar	10	生成规则的时间
Participle of interest rule	是	Varchar	10	兴趣规则的分词
Time of interest rules	是	Varchar	10	兴趣规则的时间
Weight of interest rules	否	Varchar	10	兴趣规则的权重
Number of use of interest rules	否	Varchar	10	兴趣规则的使用次数

表 4 用户与规则关联表

字段名	是否为主键	字段类型	字段大小	说明
Search Number	是	Int	11	搜索编号
Book Resources Category	否	Varchar	10	图书资源类别
Digital Library Resources	否	Varchar	10	图书数字资源
Rule number	是	Varchar	10	规则编号
Time	是	Varchar	10	时间

表 5 搜索结果分值表

字段名	是否为主键	字段类型	字段大小	说明
Rule number	是	Int	11	规则编号
Result No.	否	Varchar	20	结果编号
Page number generated by rule	否	Varchar	10	规则产生的页码
Time of Rules generated	是	Varchar	10	规则的时间
Base of rules of interestinterest rule	是	Varchar	10	兴趣规则的基数
Rating of rules of interestrules	是	Varchar	10	兴趣规则的评分
Weight of interest rules	否	Varchar	20	兴趣规则的权重
Interest Rules	否	Varchar	10	结果与兴趣相似度

2.4 系统实现效果

经过对个性化检索需求分析后, 根据系统的总体规划设计, 系统需要开发多个功能, 才能达到对图书数字资源个性化检索系统的检索效果. 主要实现的功能为: 系统检索主页、注册用户登录以及用户评分管理等.

(1) 系统检索主页

用户可以通过图书数字资源个性化检索系统中主页搜索框进行搜索图书数字资源, 也可以通过使用系统的分类的图书类型针对性的检索. 当用户输入关键词检索后, 系统会根据用户的检索历史记录, 建立个性化检索模型, 完成初次检索结果的展示, 系统以素雅的页面展示, 使得用户更加清晰看到图书数字资源数量, 实现效果如图 5 所示.

图 5 系统检索主页图

(2) 注册用户登录

图书数字资源个性化检索系统支持注册用户登录系统功能, 注册用户输入用户名和密码后, 可以进入用户中心管理后台, 登录页面实现效果如图 6 所示.

图 6 注册用户登录图

(3) 用户评分管理

用户评分管理是用于记录用户在查看某类图书数字资源的评价情况. 当某个用户给出某类图书数字资源的评分后, 作为用户兴趣模型构建因素, 影响着用户兴趣模型, 用户评分管理实现效果如图 7 所示.

图 7 用户评分管理图

2.5 系统关键步骤的实现代码

(1) 算法调用. 根据登录用户信息的情况, 结合图书数字资源的评分记录, 决定调用不同的算法实现构建个性化推荐模型, 部分代码实现如下:

```
public String recommend(){
User cUser = getCurrentUser();//获取当前登录用户
CFUtil cfUtil = new CFUtil();//实例化协同过滤推荐工具类
List<BaseModel>allScorerecords=scorerecordServic e. find(null, new Scorerecord());//获取所有评分记录
List<BaseModel>allItems= itemService. find(null, new Item());//所有项目
model = cfUtil. getDadaModel(cUser, allItems, allScorerecords);//获取用户-项目评分矩阵 List<Item> cfItemBaseUser = cfUtil. baseUser(cUser, allItems, model); //基于用户的推荐
request. setAttribute("cfItemBaseUser", cfItemBaseUser);
return "recommendSuccess";
}
```

(2) 获取数据. 根据用户的搜索请求, 系统在整个 Internet Web 上进行检索并获取符合用户需求的图书数字资源数据, 部分代码实现如下:

```
public static void dataUtil(String realPathParam, //获取数据
ItemService itemServiceParam, TypeService typeServiceParam){
realPath = realPathParam;
itemService = itemServiceParam;
typeService = typeServiceParam;
getTag();
}
```

(3) 推荐实现. 根据用户个性化推荐模型, 对获取的数据进行预处理, 最后实现推荐搜索结果, 部分代码实现如下:

```
System. out. println("推荐项目与预测评分: ");//定义推荐的项目集合
List<Item>cfItem = new ArrayList<Item>();
for(RecommendedItem ri:recommendations){//循环得到推荐项目对象
int itemid = (int) ri. getItemID();//推荐项目 id
float score = ri. getValue();//预测评分
System. out. println(ri. getItemID()+", "+score);
for(BaseModel basemodel:allItem){
Item item = (Item) basemodel;
if(item. getId()==itemid){
cfItem. add(item);
break;
}
}
}
return cfItem;
} catch (TasteException e) {
e. printStackTrace();
}
return null;
}
```

3 基于 Web 知识发现图书数字资源个性化检索系统研究

3.1 用户行为对兴趣度的影响

用户通过注册、登陆后完成相应的图书数字资源的搜索功能, 设计 3 种身份的搜索, 第 1 种是访客, 第 2 种是新用户注册后完成搜索, 第 3 种已注册用户的搜索. 用户界面所获取的信息, 能够利用 Web 知识发现, 跟踪用户对图书数字资源的兴趣度. 该技术影响下, 更加全面的用户行为信息获取与更新, 能够防止用户兴趣淘汰过快的现象产生. Web 浏览器通过后台的智能代理运行, 能够时刻监测用户对图书数字资源浏览的一切行为, 并将这些行为填加到书签, 包括摘要信息、次数信息、时间信息等, 通守返回的信息充分了解用户的兴趣度.

3.2 个性化模型的更新

图书数字资源个性化检索系统运行过程中, 用户短期兴趣、长期兴趣、新兴趣会存在较大的不同, 此时需要使用个性化的模型更新, 从而时间掌握用户的变化特征. 上述掌握用户行为的信息后, 利用 Web 知识发现后台的智能代理, 抽取相关信息特征, 智能代理具有捕捉功能, 可以将用户在 Web 网页中的操作分类, 并建立新特征词表, 从而更多好的达成各因子的筛选, 并通过滤信息, 去除用户不感兴趣的信息, 保证信息获取的有效性、准确性.

3.3 搜索结果的处理

个性化模型的更新完成后, 返回至系统后台, 依据所获得的结果与用户兴趣对应, 进再进行自动过滤, 智能加工处理信息后, 推测用户的兴趣, 给予个性化的服务. 在这一过程中的处理方法, 主要是通过个性化模型和每一个页面对应找到相似度, 经过快速反应处理后, 会将最优的搜索结果返回至用户页面, 用户会直接获得所需信息.

4 结论

在互联网时代, 大量图书数字资源信息在网络中充斥, 用户对于图书数字资源的需求也越来越多, 然而面临海量的图书数字资源信息, 如何快速定位感兴趣的资源, 已经成为当前用户的核心需求. 由此可见, 图书数字资源个性化检索系统的设计是一种必然发展趋势. 传统图书资源的基础之上, 打造个性化的搜索引擎, 能够更好的挖掘用户需求, 匹配用户搜索关键词, 优化搜索结果, 快速返回用户, 最终使用户获得符合自身兴趣的结果. 本次研究中基于 Web 知识发现和与智能代理等技术相结合, 深入挖掘客户的兴趣信息, 利用后台智能代理技术获取用户浏览过程中一切行为, 通过所获取的用户兴趣信息, 利用个性化技术建立个性化模型. 在个性化模型中, 过滤用户不感兴趣的搜索结果, 并通过优化搜索结果, 快速返回至用户, 从而真正实现用户所需信息的获取.

参考文献

1 王新才. 知识发现系统与通用学术搜索引擎文献资源比较研究——以超星发现和百度学术为例. 福建论坛·人文社会科学版, 2018, (4): 164–172.

2 李洁, 毕强, 许鹏程, 等. 数字图书馆知识发现的数据驱动机制及绩效优化研究. 图书情报工作, 2019, 63(3): 5–13.

3 Aly M, Pandey S, Josifovski V, *et al*. Towards a robust modeling of temporal interest change patterns for behavioral targeting. Proceedings of the 22nd international conference on World Wide Web. Rio de Janeiro, Brazil. 2013. 71–82.

4 杨瑞峰. Web 上基于文本挖掘的个性化检索系统的设计与实现 [硕士学位论文]. 成都: 电子科技大学, 2004.

5 曾艳. 图书馆数字资源共享与服务理论与实践研究——评《数字图书馆特色资源共享与服务研究》. 图书馆工作与研究, 2019, (2): 2.

6 朱江, 任晓亚, 姜恩波, 等. 研究图书馆数字资源建设的转型与发展——以中国科学院文献情报系统为例. 图书情报工作, 2019, 63(4): 47–53.

7 王陆, 彭功, 马如霞, 等. 大数据知识发现的教师成长行为路径. 电化教育研究, 2019, 40(1): 95–103.

8 管皓, 秦小林, 饶永生. 动态数学数字资源开放平台的研究与设计. 哈尔滨工业大学学报, 2019, 51(5): 14–22. [doi: 10.11918/j.issn.0367.6234.201712127]

9 Johnson JA, Liu MC, Chen H. Unification of knowledge discovery and data mining using rough sets approach in a real-world application. Proceedings of the Second International Conference on Rough Sets and Current Trends in Computing. Banff, AB, Canada, 2000. 330–337.

10 邵丝媞. 基于空间信息支持的图书馆个性化资源集成系统设计. 现代电子技术, 2019, 42(18): 112–115, 119.

11 李默. 数字图书馆个性化移动视觉搜索机制研究. 图书馆理论与实践, 2019, (2): 107–112.

12 周谦豪, 戴泽钒, 朱奕帆, 等. inBooks 数字人文工具的设计与实现——基于上海图书馆开放数据的微信小程序. 图书馆杂志, 2019, 38(2): 41–48, 68.

13 Hayes-Roth B. Putting intelligent characters to work. AI Magazine, 2008, 29(2): 43–48.

14 William JF, Gregory PS, Christopher JM. Knowledge discovery in databases: An overview. AI Magazine Knowledge Discovery, 1992, 13(3): 213–228.