



北京交通大学

基于大数据的智能推荐系统设计与实现

数据收集

张迪

dizhang@bjtu.edu.cn

项目需要的基本数据

- 业务数据

功能	所属表	所属页面
登录	user	登录页面
图书列表展示	Book_info、book_category	主页面
图书详情显示	Book_info、book_desc	图书详情页面
图书打分（0~10分）	Bx_book_ratings	图书详情页面、购物车页面
生成订单和订单明细	orders、order_detail	购物车页面
模拟支付	orders、order_shipping	购物车页面、支付页面

项目需要的基本数据

- 日志数据

数据	说明	
网站访问日志	用于用户点击流日志分析	1、 coolshell_20140212.log（演示用） 2、 apache_simple.log(样例日志数据) 3、 本项目Web端产生的Apache访问日志

数据库设计

- 基本表格

表名	功能	备注
users	用户表	
book_info	图书信息表	
bx_book_ratings	用户图书评分表	
book_category	图书类别表	
book_desc	图书描述表	
orders	订单表	
order_detail	订单详情表	

















数据库设计

- 基本表格

表名	功能	备注
order_shipping	订单邮寄表	
privilege	权限表	
reply	回复表	
role	角色表	
role_privilege	权限角色对应表	
store	店铺表	
user_role	用户角色表	
comment	评论表	
global_parameter	全局参数表	

项目数据

- 项目数据为csv格式

 book_category.csv	2019/5/19 21:15	Microsoft Excel ...	1 KB
 book_desc.csv	2019/5/19 21:15	Microsoft Excel ...	26 KB
 book_info.csv	2019/5/19 21:16	Microsoft Excel ...	97,518 KB
 bx_book_ratings.csv	2019/5/19 21:16	Microsoft Excel ...	24,775 KB
 comment.csv	2019/5/19 21:16	Microsoft Excel ...	0 KB
 global_parameter.csv	2019/5/19 21:16	Microsoft Excel ...	1 KB
 order_detail.csv	2019/5/19 21:16	Microsoft Excel ...	301 KB
 order_shipping.csv	2019/5/19 21:16	Microsoft Excel ...	1 KB
 orders.csv	2019/5/19 21:16	Microsoft Excel ...	200 KB
 privilege.csv	2019/5/19 21:16	Microsoft Excel ...	4 KB
 reply.csv	2019/5/19 21:16	Microsoft Excel ...	0 KB
 role.csv	2019/5/19 21:16	Microsoft Excel ...	1 KB
 role_privilege.csv	2019/5/19 21:16	Microsoft Excel ...	7 KB
 store.csv	2019/5/19 21:16	Microsoft Excel ...	1 KB
 user.csv	2019/5/19 21:16	Microsoft Excel ...	31,767 KB
 user_role.csv	2019/5/19 21:16	Microsoft Excel ...	1 KB

爬取数据与项目数据的合并

- 爬虫完成情况
 - 爬取了哪些图书信息？
- 根据项目数据表的信息修改爬虫
 - 修改编写的爬虫使得能够抓取项目信息表所需要的图书信息数据
- 用自己爬取的图书信息替换项目数据中的图书信息
 - 目的：订单信息相同，但是图书信息不同

日志数据

- 项目所需的访问日志的格式为标准的Apache日志

1	10.178.123.55	远端主机
2	-	远端登录名
3	-	远程用户名
4	[11/Dec/2018:10:00:32 +0800]	服务器接收时间
5	"GET /__utm.gif HTTP/1.1"	请求的第一行(请求的路径)
6	200	最后请求的状态
7	35	以CLF格式显示的除HTTP头以外传送的字节数
8	"http://easternmiles.ceair.com/flight/index.html"	上一个访问页面
9	"Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/ 537.36 (KHTML, like Gecko) Chrome/31.0.1650.63 Safari/537.36"	用户浏览器信息
10	"BIGipServermu_122.119.122.14=192575345.20480.000;Webtrends= 120.196.145.58.1386724976245806; "	Cookie信息
11	1482	接收的字节数, 包括请求头的数据, 且不能为零
12	352	发送的字节数, 包括请求头的数据, 且不能为零
13	-	%(X-Forwarded-For)i
14	easternmiles.ceair.com	访问主机地址 (域名)
15	749	服务器处理本请求所用的时间, 以微秒为单位

日志数据

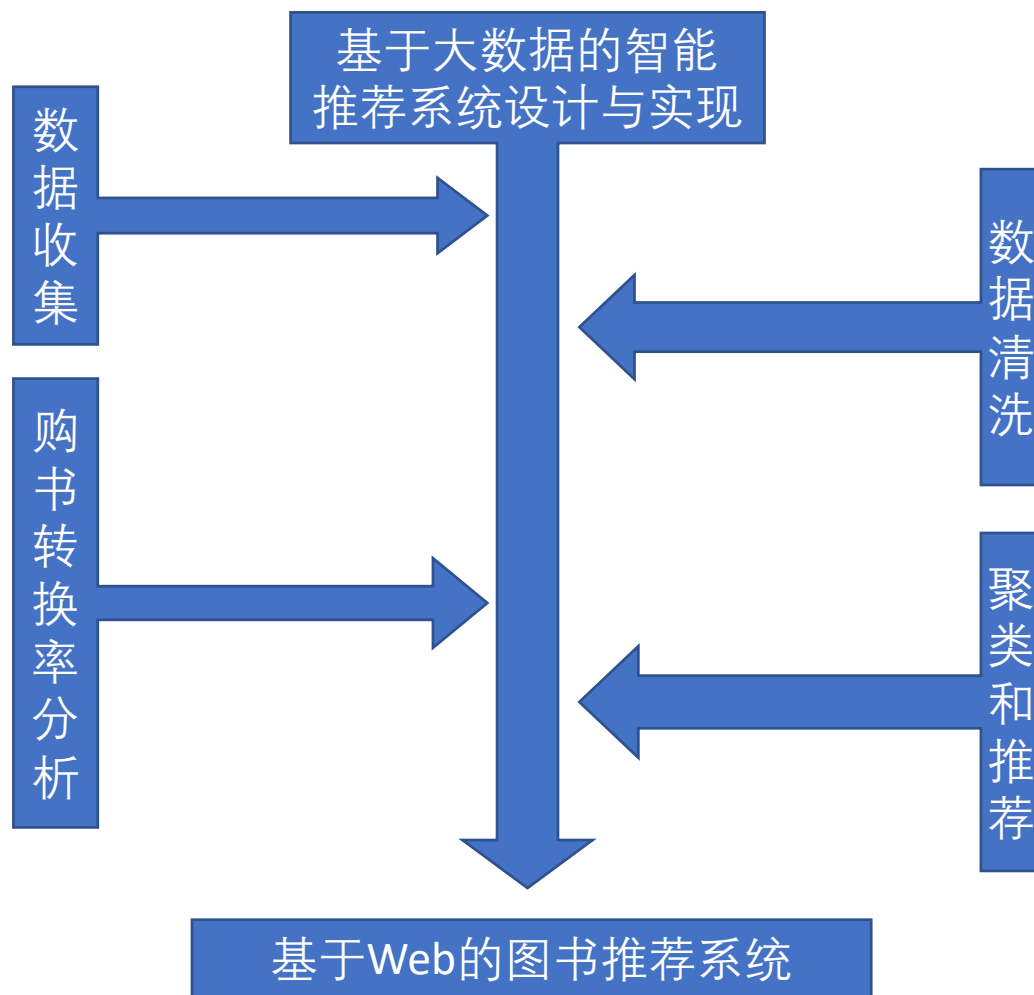
- 数据来源是由本项目的Web端程序运行所产生的访问日志
- 修改Tomcat日志格式，使Tomcat服务器产生的日志格式符合上面表格中的格式要求：
 - 如果使用的是纯Apache服务器，那么访问日志在Apache根目录的conf文件夹内的httpd.conf文件中配置
 - 如果使用的是Apache tomcat服务器，那么访问日志在Apache根目录的conf文件夹内的server.xml文件中配置
 - Apache Tomcat访问日志的最终格式设置为：`pattern="%h %l %u %t
"%r" %s %b "%{Referer}i" "%{User-Agent}i"
"%{uuid}c;%{userId}c;%{st}c;" %l %O %{X-Forwarded-For}i %v %D"`

相关任务

- 完成大数据开发环境的配置和搭建
- 根据给出的指导手册完善需求分析
- 爬取图书数据并与所给的数据中的其他信息合并
- 配置Web服务器获取访问日志数据

后续任务

- 数据分析处理与图书推荐系统开发并行



特别注意

- 环境配置完成后注意事项
 - 关机流程
 - 先在master上执行stop-all.sh
 - 然后再关机
 - 开机流程
 - 先开机
 - 然后，在master上执行start-all.sh



北京交通大学

Thank you!

Q & A