

目 录



01. 背景与意义

02. 方法与思路

03. 关键技术及成果

04. 总结与展望

[PART 01]

[背景与意义]





高校图书馆扮演着传播知识与文化的圣地，也是高校学生的学习课堂。高校图书馆承载着巨量的图书资源，阅读推广不仅有利于促进学生的学习氛围还有助于提高图书馆资源的利用率。而由于缺乏对读者的阅读兴趣和习惯的分析，阅读推广的效果往往不达标。因此，利用图书借阅数据来分析借阅行为，可以为阅读推广提供有益的决策和帮助。

随着大数据技术的成熟应用，个性化推荐系统在许多领域大放异彩，而高校图书馆拥有者庞大的用户群体和图书资源，积累了大量的用户行为数据。而对于百万级甚至千万级馆藏量的高校图书馆来说，传统的阅读推广模式已经无法适应于新时代的需求。



随着大数据技术的不断发展和逐渐成熟，个性化推荐系统已经在电商、新媒体、短视频平台等领域大放异彩。通过对用户行为数据分析，挖掘用户潜在的兴趣并据此个性化的推荐物品。相较于传统的摆摊模式，该方法更具有主动性和确定性。

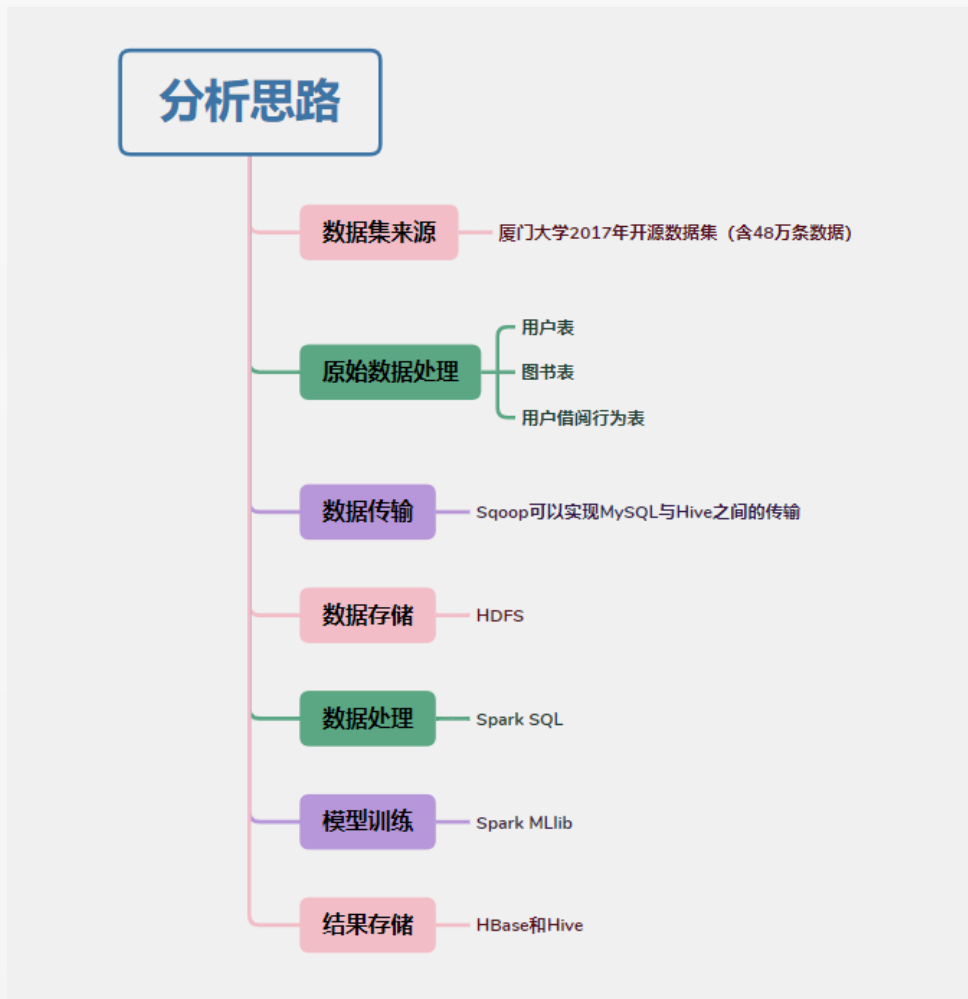
面对高校图书馆积累的大量借阅行为数据，采用数据挖掘和大数据技术挖掘兴趣和习惯。处理高校图书馆积累的注册的用户数据、图书数据和用户在高校图书检索系统中的与借阅相关的行为数据，进而构建画像结合成熟的推荐算法和机器学习优化算法对读者阅读需求、阅读兴趣、阅读习惯等进行预测，告别传统的粗放型的阅读推广方式，从而转向细致化。

[PART 02]

[方法与思路]



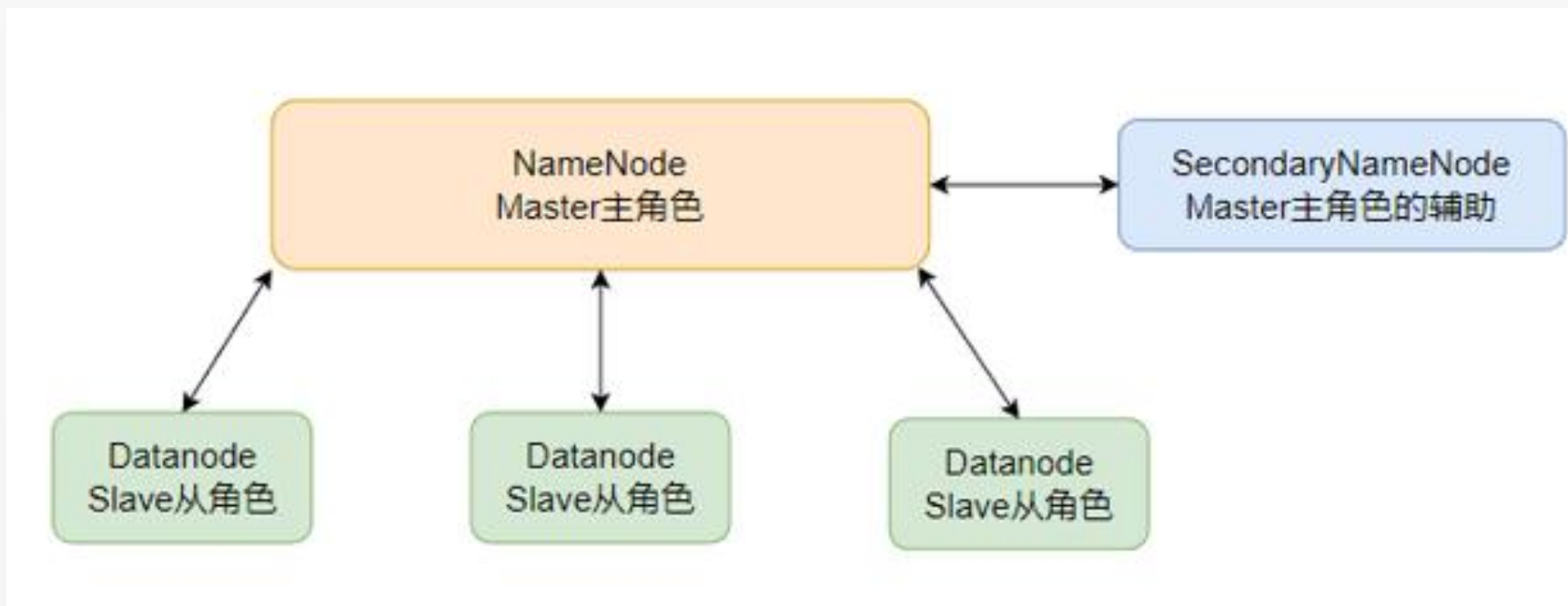
通过分析选题内容、阅读相关论文文献并调研相关技术，确定了如下实现思路。



Hadoop分布式计算框架

随着大数据时代的到来，用户规模日益剧增，存储数据呈几何时增长，传统的数据存储方案已经落后于时代，因此分布式存储和计算框架Hadoop应运而生。

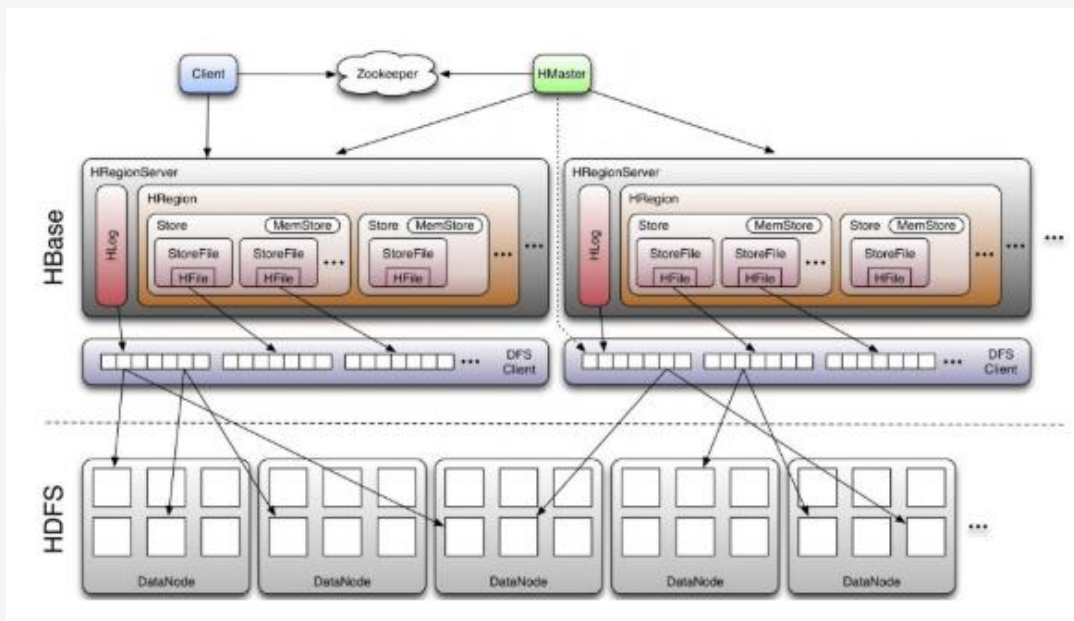
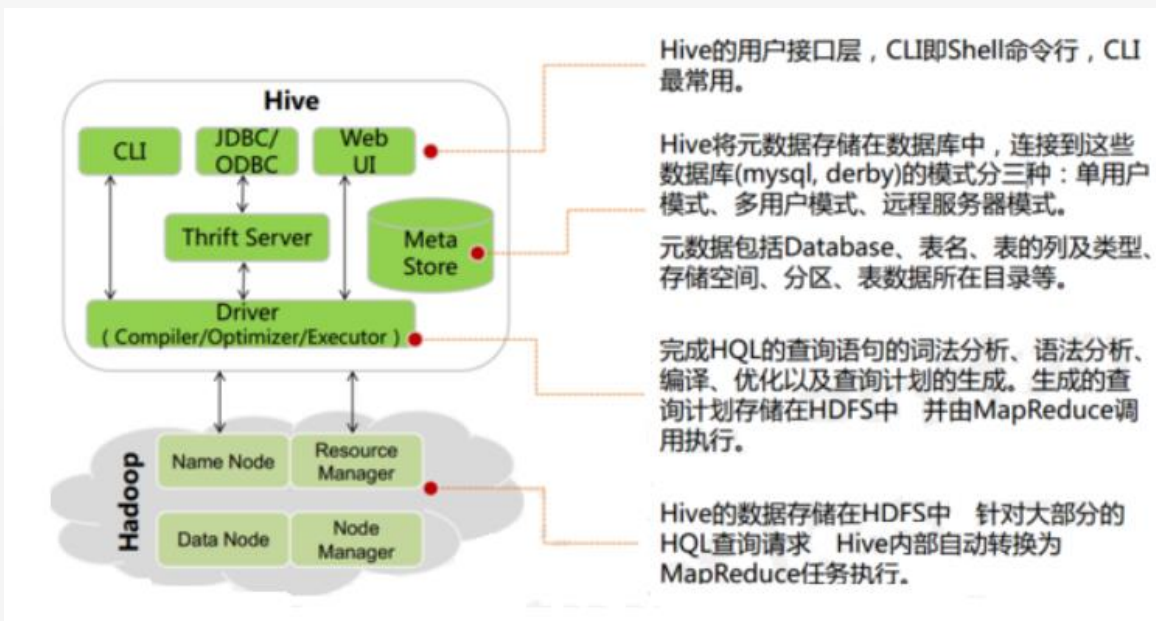
其中的HDFS早已成为当今最流行的工业级数据仓库之一，它的核心功能是提供高效稳定的数据存储服务。HDFS具备良好的负载均衡和并行处理能力，能够扩展到以PB级别的数据规模存储。



Hive和Hbase

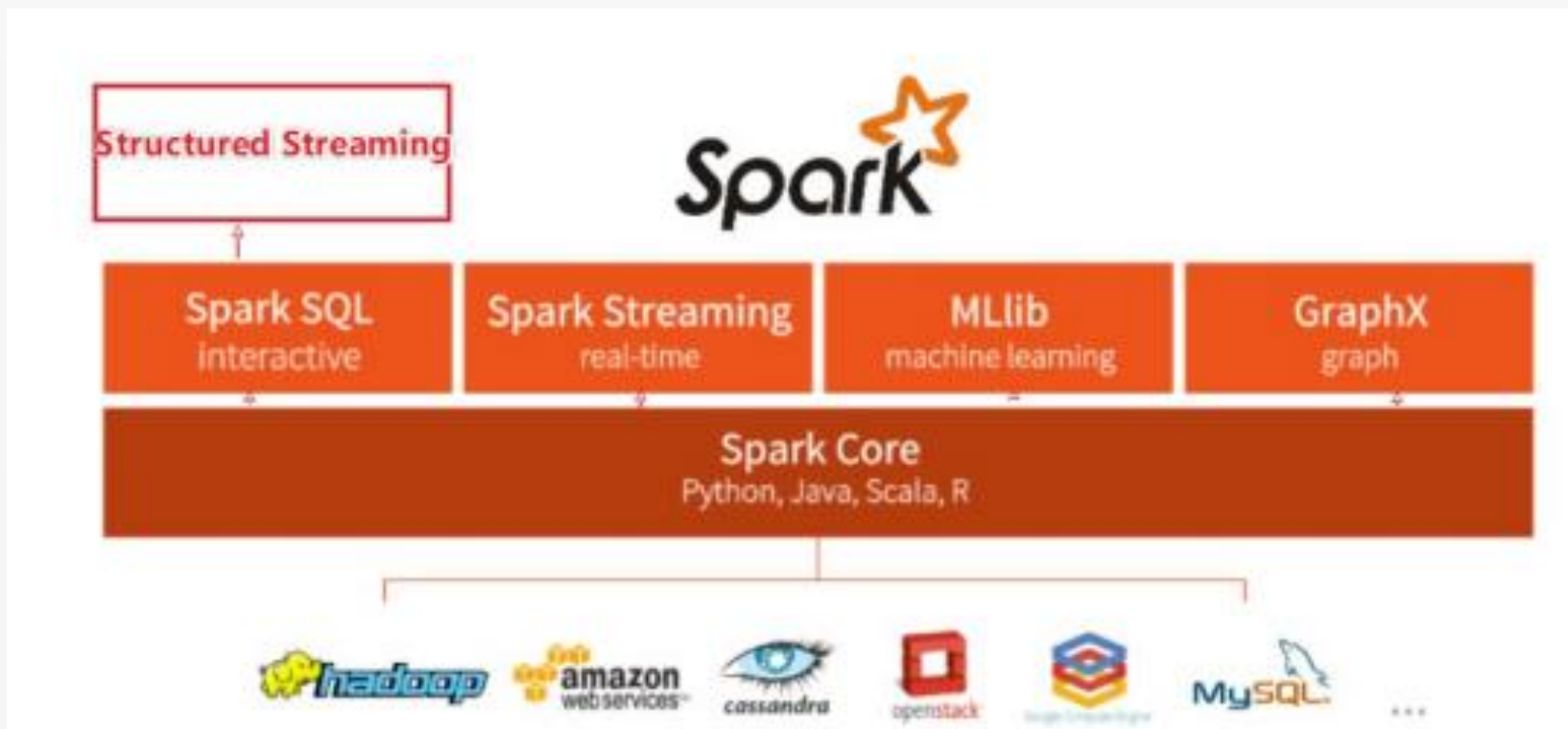
Hive是一个基于Hadoop的开源数据仓库系统，它能够将结构化的数据映射为数据库表格，并将底层数据存储在HDFS分布式文件系统中。

HBase由大量的服务器节点组成集群，每个节点都可以存储和处理数据，这些节点可以扩展到数千台服务器，以实现水平扩展。HBase使用HDFS作为底层存储系统。



计算引擎Spark

Spark是一款基于内存式的高性能计算引擎，它通过将数据存储在内存中，将计算速度提高了数倍。相较于Hadoop MapReduce，Spark更快，更高效利用内存，支持更多的复杂操作和提供简单易用的编程接口。此外，Spark机器学习组件也受到广泛的关注，通过支持分布式迭代计算，可以快速训练大规模的机器学习模型。Spark与Hadoop生态有完美的集成，将Hadoop中的各种组件（如HDFS和YARN）作为底层存储和调度平台，并提供各种API，便于用户快速开发和管理Spark应用程序。



基于协同过滤的算法

基于用户的协同过滤算法

UBCF算法的思想为：倘若两个用户历史上有很多的行为和喜好相似，则在某些方面他们的兴趣可能也相似，故可以根据用户之间的相似度来预测一个用户对物品的喜好程度。这里的相似度可以使用如余弦相似度、皮尔逊相关系数等距离度量方法来计算。UBCF算法的优点是简单易懂，推荐结果具有可解释性，但其缺点也显而易见，即当用户和物品的数量很大时，计算相似度会非常耗时，并且由于用户的爆炸式增长，导致很多用户仅仅交互了几个物品，难以找到足够相似的其他用户，因此准确度会受到很大影响。

基于物品的协同过滤算法

IBCF算法的思想是利用相似度来进行推荐。相似度的计算不再是基于用户，而是基于物品。即对于一个物品A，首先计算其与其他物品的相似度，然后等权重地考虑用户历史上与A类似物品的兴趣度，作为用户对于A物品的兴趣度的预测值。与UBCF相比，IBCF算法的优点是计算量更小，而且相对容易实现。但因为有些物品的交互数量很小，所以难以找到足够相似的物品，导致推荐效果不佳。

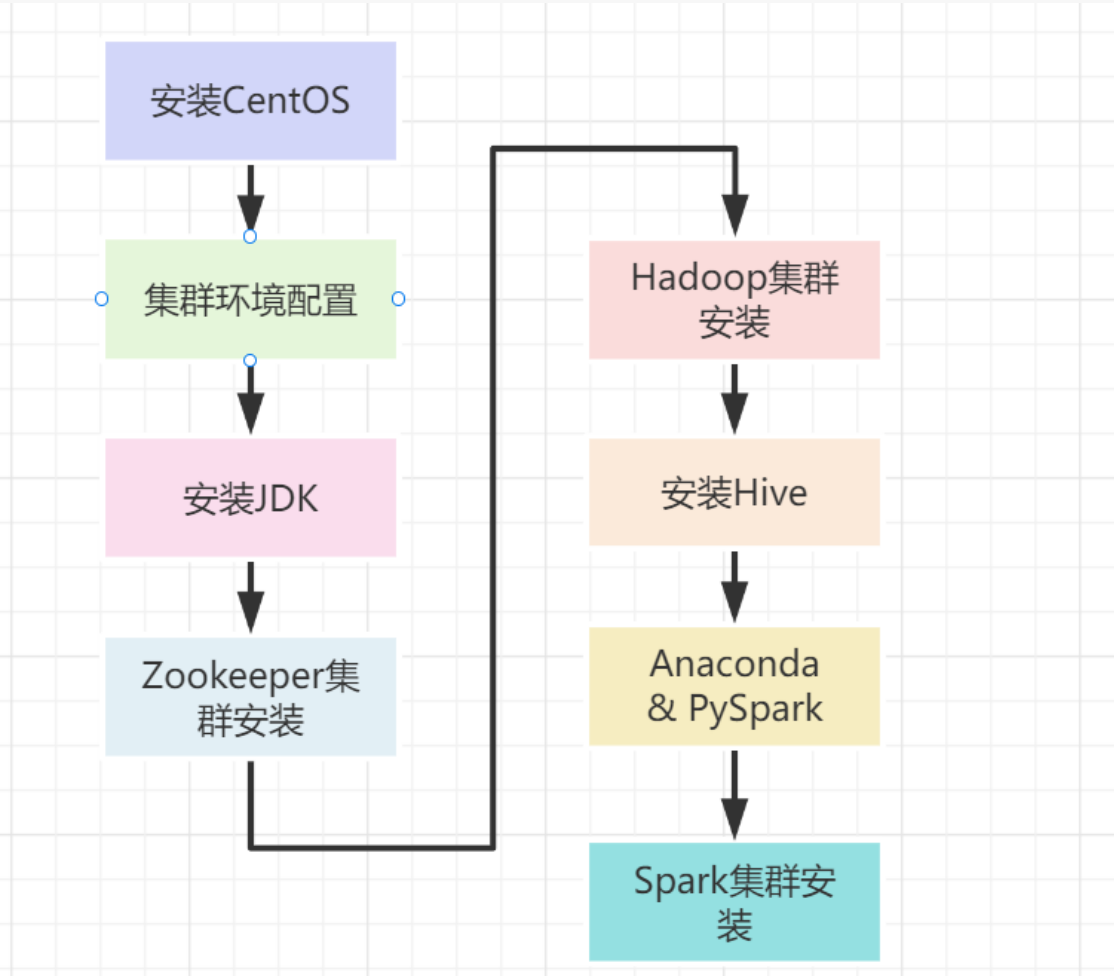


[PART 03]

关键技术及成果



大数据集群的搭建流程



大数据集群的主要配置

表 5-1 集群配置信息

节点名称	CPU 核心数	内存（GB）	系统
ubuntu-master	2	4	Ubuntu22.04
ubuntu-slave1	2	4	Ubuntu22.04
ubuntu-slave2	2	4	Ubuntu22.04

服务器 IP	主机名	myid 的值
192.168.248.147	Hadoop-master	1
192.168.248.148	Hadoop-slave1	2
192.168.248.149	Hadoop-slave2	3

节点	主节点(master)	从节点(worker)	历史服务(history server)
Hadoop-master	是	是	是
Hadoop-slave1	否	是	否
Hadoop-slave2	否	是	否

原始数据处理

数据集来源于选自厦门大学2017年图书馆借阅数据集，文件大小为129M，文件格式为.csv文件，总共包含487038条借阅数据。包含以下信息：

用户ID（READER_ID），
性别（READER_SEX），
所在学院（READER_DEPT），
所属年级（READER_GRADE），
借阅时间（LEND_DATE），
归还时间（RET_DATE），
续借次数（RENEW_TIMES），
馆藏位置（LOCATION_NAME），
所借书籍名称（M_TITLE），
索书号（M_CALL_NO），
条码号（M_ISBN），
作者（M_AUTHOR），
出版社（M_PUBLISHER），
年卷期（M_PUB_YEAR），
文档类型（DOC_TYPE_NAM）。

2017.csv - Excel																
文件 开始 插入 绘图 页面布局 公式 数据 审阅 视图 帮助 操作说明搜索																
剪贴板 剪贴 复制 格式刷 剪贴板 字体 等线 11 常规 条件格式 套用 单元格样式 插入 删除 格式 自动求和 填充 清除 排序和筛选 查找和选择 编辑																
A1 X READER_ID																
A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	READER_ID	READER_SEX	READER_DEPT	READER_GRADE	READER_TYPE	LEND_DATE	RET_DATE	RENEW_TIMES	LOCATION_NAME	M_TITLE	M_CALL_NO	M_ISBN	M_AUTHOR	M_PUBLISHER	M_PUB_YEAR	DOC_TYPE_NAM
2	015705502677	M	嘉庚学院	2013	嘉庚本科	2017-06-27	2016-09-26	17	0 嘉庚馆藏 (漳州校区)	华夏意匠:中国古	TU2/106.1	978-7-5618-4	李允铎著	天津大学出版社	2014	中文图书
3	4941A18DD4D	M	航空航天大学	2015	本科生	2017-06-27	2016-07-20	20	0 总馆基本书库	大学语文与写作	H1/482	978-7-302-33	主编吴满珍	清华大学出版社	2013	中文图书
4	4941A18DD4D	M	航空航天大学	2015	本科生	2017-06-27	2016-08-28	14	0 基本书库-新书区	英国现代转手	D756.19/818	978-7-108-04	刘成著	生活·读书·新知三联	2013	中文图书
5	4941A18DD4D	M	航空航天大学	2015	本科生	2017-06-27	2016-07-20	20	0 基本书库-新书区	军情五处与谍	D756.136/29	978-7-213-04	(英)戈登·托马	浙江人民出版社	2012	中文图书
6	3E209E279461	F	建筑与土木工程	2016	硕士生	2017-06-27	2016-07-18	9	0 建筑土木工程资料	建造设计手册	TU206/229.2	978-7-5537-6	(德)本杰明·胡	江苏科学技术出版	2016	中文图书
7	3E209E279461	F	建筑与土木工程	2016	硕士生	2017-06-27	2016-07-18	9	0 建筑土木工程资料	2010 Revit杯	TU206/647.1	978-7-112-13	全国高等学校	中国建筑工业出版	2011	中文图书
8	3E209E279461	F	建筑与土木工程	2016	硕士生	2017-06-27	2016-07-18	9	0 建筑土木工程资料	2014 AUTOD	TU206/647.1	978-7-112-19	全国高等学校	中国建筑工业出版	2016	中文图书
9	3E209E279461	F	建筑与土木工程	2016	硕士生	2017-06-27	2016-06-29	15	0 建筑土木工程资料	地区基础设施	TU206/180.2	978-7-5611-5	日本株式会社	大连理工大学出版	2010	中文图书
10	F85D456598E4	F	外文学院	2015	硕士生	2017-06-27	2016-10-17	16	1 总馆外文书库	日本伦理思想	B82-093.13/4	4000001563X	4和辻哲郎著	岩波書店	1979	日文图书
11	0649014A736E	M	软件学院	2016	本科生	2017-06-27	2016-10-27	9	1 信息工程分馆	Java面向对象	TP312JA/022	978-7-5635-2	张桂珠, 张平,	北京邮电大学出版	2010	中文图书
12	1398E889F3CB	M	艺术学院	2015	本科生	2017-06-27	2016-09-21	16	0 基本书库-新书区	飞越疯人院	I712.45/113.2	978-7-229-07	(美)肯·克西著	重庆出版社	2015	中文图书
13	038A7ABB332	M	建筑与土木工程	2015	硕士生	2017-06-27	2016-09-20	14	0 建筑土木工程资料	道·设计·建筑	TU2/672	978-7-111-39	徐守珩著	机械工业出版社	2013	中文图书
14	038A7ABB332	M	建筑与土木工程	2015	硕士生	2017-06-27	2016-09-20	14	0 建筑土木工程资料	飘渺余蕴天	TU-885/181	978-7-112-13	编著王谢燕	中国建筑工业出版	2011	中文图书
15	8EAC8474349E	M	化学化工学院	2014	本科生	2017-06-27	2016-10-22	12	1 总馆基本书库	温故一九四二	I247/811.2	978-7-02-006	刘震云,	人民文学出版社	2009	中文图书
16	8EAC8474349E	M	化学化工学院	2014	本科生	2017-06-27	2016-10-20	12	1 总馆基本书库	沈从文小说	I246.7/867.1	978-7-02-003645	- 沈从文著	人民文学出版社	1982	中文图书
17	8EAC8474349E	M	化学化工学院	2014	本科生	2017-06-27	2016-10-22	12	1 总馆基本书库	淮比淮傻多	I247.5/182.1	978-7-5302-1	王朔著	北京十月文艺出版	2012	中文图书
18	8EAC8474349E	M	化学化工学院	2014	本科生	2017-06-27	2016-10-20	12	1 总馆基本书库	沈从文小说	I246.7/867.1	978-7-02-003645	- 沈从文著	人民文学出版社	1982	中文图书
19	8EAC8474349E	M	化学化工学院	2014	本科生	2017-06-27	2016-08-29	11	0 基本书库-新书区	美洲豹阳光	I546.45/360.2	978-7-5447-5	伊塔洛·卡尔维	译林出版社	2015	中文图书
20	8EAC8474349E	M	化学化工学院	2014	本科生	2017-06-27	2016-08-29	11	0 基本书库-新书区	圣约翰之路	I546.45/360.1	978-7-5447-5	(意)伊塔洛·卡	译林出版社	2015	中文图书
21	8EAC8474349E	M	化学化工学院	2014	本科生	2017-06-27	2016-10-22	12	1 总馆基本书库	爱向虚空茫茫	I247.5/178.5	978-7-5321-4	王安忆	上海文艺出版社	2013	中文图书
22	F425D033F97C	M	台湾研究院		教职工	2017-06-27	2016-07-07	10	0 区域研究资料中心	财訊, 264期	(F8-55/806/2)	1729-3758	曾熾卿	財訊雜誌社		中文图书
23	F425D033F97C	M	台湾研究院		教职工	2017-06-27	2016-07-07	10	0 区域研究资料中心	财訊, 263期	(F8-55/806/2)	1729-3758	曾熾卿	財訊雜誌社		中文图书

要想实现以高校图书馆用户借阅行为数据分析为基础的推荐系统的建立，需要在现有的用户行为数据中筛选出对用户借阅行为偏好和习惯体现力度较强的数据，增强训练模型的精度和多维度。因此针对上述数据集，将该数据集经过去重，拆分操作之后得到三个文件，分别为用户信息表（**reader**），图书信息表（**book**）和行为表（**reader_action**）。

▼	reader
▼	columns 6
	id int (auto increment)
	reader_no varchar(255)
	reader_sex varchar(2)
	reader_dept varchar(255)
	reader_grade varchar(255)
	reader_type varchar(255)
▼	keys 1
	PRIMARY (id)

▼	book
▼	columns 10
	id int (auto increment)
	location_name varchar(255)
	m_title varchar(255)
	call_no varchar(255)
	m_call_no varchar(255)
	m_isbn varchar(255)
	m_author varchar(255)
	m_publisher varchar(255)
	m_pub_year varchar(255)
	doc_type_name varchar(255)
▼	keys 1
	PRIMARY (id)

▼	reader_action
▼	columns 9
	id int (auto increment)
	reader_no varchar(255)
	reader_id bigint
	book_id bigint
	call_no varchar(255)
	lend_date varchar(255)
	ret_date varchar(255)
	renew_times int
	m_title varchar(255)
▼	keys 1
	PRIMARY (id)

用户信息表（reader）包含ID（对原有的读者ID进行编码，原数据存在冗长的缺陷不便于数据处理），原数据中的读者ID，性别，所属学院，入学年份以及身份类型。

reader						
	id	reader_id	reader_sex	reader_dept	reader_grade	reader_type
0	1	015705502677A5A8572D479D4ABE55F5	M	嘉庚学院	2013.0	嘉庚本科
1	2	4941A18DD4DB75730FA67B47B22BD959	M	航空航天学院	2015.0	本科生
2	3	3E209E279461206602794D1837CEA3D1	F	建筑与土木工程学院	2016.0	硕士生
3	4	F85D45659BE41B9F2F6BB74173EE5E41	F	外文学院	2015.0	硕士生
4	5	0649014A736EB7150AA2A94B1E6108C7	M	软件学院	2016.0	本科生
...
39355	39356	064EDDBCA8257E111EC5728A94EBE097	F	财务管理与会计研究院	2017.0	硕士生
39356	39357	E897458AA0C7A7409E460C1CC3DF7EE2	F	经济研究所	2017.0	交流生
39357	39358	B7181898CD58AD6EF95F2C94505BCF69	M	经济学院	2016.0	硕士生
39358	39359	E2F2F73D58A0E117A2B19B3A22994CBE	F	嘉庚学院	2014.0	嘉庚本科
39359	39360	E371636EA720CFCD010EE65ACD4886D0	M	南洋研究院	2017.0	硕士生

35903 rows × 6 columns

原始数据处理

图书信息表（book）包含id编码，图书所在藏馆，书名，索书号，条码号，作者，出版社，年卷期和文档类型，其中的藏馆位置，书名，索书号第一层级（例如：TP311.5），作者信息和文档类型可以作为后续的标志。

```
In [20]: book = pd.read_csv("data/book.csv", usecols=range(0, 9)).dropna()
book
```

Out[20]:

	id	location_name	m_title	m_call_no	m_isbn	m_author	m_publisher	m_pub_year	doc_type_name
0	1	信息工程分馆	软件工程与计算.卷二,软件开发的技术基础.Volume II,Fundamentals of ...	TP311.5/088.21/(2)	978-7-111-40750-8	骆斌主编	机械工业出版社	2012	中文图书
1	2	信息工程分馆	软件工程与计算.卷一,软件开发的编程基础.Volume I,Programming fund...	TP311.5/088.21/(1)	978-7-111-40697-6	骆斌主编	机械工业出版社	2012	中文图书
2	3	基本书库-新书区	中国汽车物流发展报告.2015	F407.471/682/(2015)	978-7-5047-5905-4	China automotive logistics association of CFLP	中国财富出版社	2015	中文图书
3	4	基本书库-新书区	流程工业多品种成批生产计划与调度	F406.2/722	978-7-122-18993-6	唐琦著	化学工业出版社	2014	中文图书
4	5	翔安分馆	战国纵横:鬼谷子的局.4,点化二子, 苏秦张仪舌战群雄	I247.53/769.12/(4)	978-7-5399-5656-5	寒川子著	江苏文艺出版社	2012	中文图书
...
197360	197361	总馆基本书库	网络光芒.Ⅱ,互联网的价值与潜质	F276.6/124.5/(2)	978-7-111-31535-3	2010中国互联网大会组委会, 中国网民文化节组委会编	机械工业出版社	2010	中文图书

用户行为表（reader_action）包含用户id，借阅时间，归还时间，续借次数，书名，索书号和条码号。而用户行为表作为用户-行为-图书结构是构建用户画像的核心数据，也是连接用户ID与图书ID的桥梁（reader_id和book_id）。

对象	book @library (ubuntu-master) - 表		reader @library (ubuntu-master) - 表		reader_action @library (ubuntu-mas...			
开始事务	文本	筛选	排序	导入	导出			
id	reader_no	reader_id	book_id	call_no	lend_date	ret_date	renew_times	m_title
	1 4941A18DD4DB75730	1630	780	H1	2017-06-27 16:26:32	2017-07-20 20:52:36	0	大学语文与实用写作
	2 4941A18DD4DB75730	1630	614	D756.19	2017-06-27 16:26:47	2017-08-28 14:43:50	0	英国现代转型与工党重
	3 8EAC847434988C91D	1769	616	I246.7	2017-06-27 16:38:32	2017-10-20 12:09:18	1	沈从文小说选
	4 8EAC847434988C91D	1769	628	I247.5	2017-06-27 16:38:40	2017-10-22 12:49:00	1	谁比谁傻多少
	5 8EAC847434988C91D	1769	616	I246.7	2017-06-27 16:38:50	2017-10-20 12:09:08	1	沈从文小说选
	6 8EAC847434988C91D	1769	629	I546.45	2017-06-27 16:38:58	2017-08-29 11:41:30	0	美洲豹阳光下
	7 8EAC847434988C91D	1769	633	I546.45	2017-06-27 16:39:06	2017-08-29 11:41:28	0	圣约翰之路
	8 8EAC847434988C91D	1769	636	I247.5	2017-06-27 16:39:13	2017-10-22 12:49:08	1	爱向虚空茫然中
	9 3BE8A092EAA79DC3E	4580	640	G210	2017-06-27 16:39:46	2017-09-21 13:50:37	0	语用学视角下的新闻转
	10 F475DF393E18A01C85	1780	643	TU984.11	2017-06-27 16:47:52	2017-09-28 16:06:35	0	交往与空间
	11 181DCDEE6FB70CC53E	1781	16090	F830.9	2017-06-27 16:47:55	2017-07-05 07:50:14	0	《期权、期货及其他衍
	12 181DCDEE6FB70CC53E	1781	18309	F830.9	2017-06-27 16:47:57	2017-07-05 07:50:11	0	期权、期货及其他衍生
	13 C838910D1C327F0052	1782	18310	B5	2017-06-27 16:48:32	2017-09-18 18:26:55	0	简明现代西方哲学
	14 49B77DF02A39A9CE3E	1784	18311	TP274	2017-06-27 16:50:07	2017-09-24 15:08:15	0	Hadoop应用开发技术
	15 A4802330289AB78EB7	1785	18312	H31	2017-06-27 16:50:23	2017-09-11 09:16:33	0	STEP BY STEP日常英语
	16 A4802330289AB78EB7	1785	18313	H31	2017-06-27 16:50:33	2017-09-11 09:16:28	0	孔雀女的英语心经
	17 246E0B3CB0974E089E	1786	18314	I267.1	2017-06-27 16:51:02	2017-06-30 15:55:50	0	活着活着就老了
	18 D933CBCFEC31D60D8	1788	18316	I217.2	2017-06-27 16:52:53	2017-07-21 12:04:27	0	汪曾祺自选集
	19 71E4917ABE92621161	1789	18318	H314	2017-06-27 16:54:10	2017-09-18 15:03:27	0	新编大学英语语法
	20 89B524D3F065A28E41	1790	18320	J524.3-39	2017-06-27 16:54:57	2017-09-25 10:01:40	0	Photoshop CS5平面
	21 0A7764D672798DE3B6	1791	18324	J905.313	2017-06-27 16:56:23	2017-09-12 17:44:50	0	日本电影
	22 3DC258199F25B53BD6	814	18326	D913.04	2017-06-27 16:56:31	2017-09-25 09:14:42	0	民法的自然法学基础
	23 8635274144E204FF60	319	18328	D913.404	2017-06-27 16:56:39	2017-09-25 09:14:30	0	著作权的宪法之维

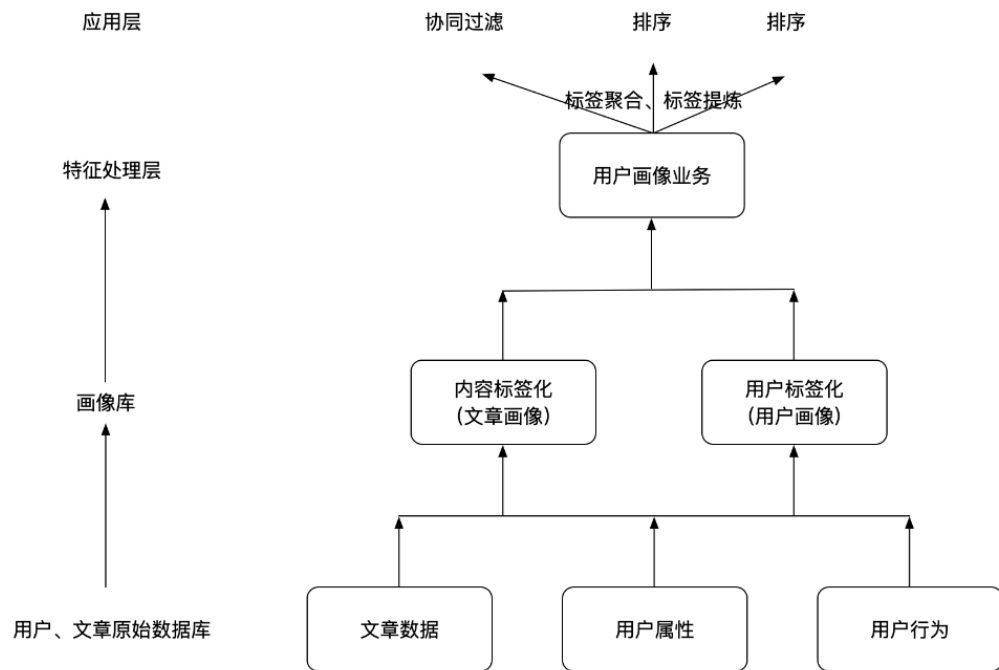
图书画像构建:

- 原始数据处理
- 使用spark完成图书Tfidf值计算
- 使用spark完成图书TextRank值计算
- 使用spark完成图书画像结果值计算与存储

图书画像，就是给每本图书定义一些词。主题词与关键词最大的区别就是主题词经过了规范化处理。

关键词：图书数据中权重高的词，使用TEXTRANK计算出的结果TOPK个词以及权重

主题词：图书中出现的同义词，计算结果出现次数高的词，TEXTRANK的TOPK词 与 ITFDF计算的TOPK个词的交集



1、原始图书表数据合并得到图书所有的词语句信息

通过spark sql 来进行操作，合并信息，由于每次调试运行spark时间较长，使用jupyter notebook进行开发可以保存一些临时变量

2、对图书进行Tf-idf计算

TFIDF模型的训练步骤：

读取所有（N本）图书数据

图书数据进行分词处理, 得到分词结果（jieba词典）

TFIDF模型训练保存，spark使用count与idf进行计算

TFIDF计算方案：

- 先计算分词之后的每本图书的词频，得到IF模型
- 然后根据词频计算IDF以及词，得到IDF模型
- 利用模型计算N本图书数据的TFIDF值

```
+-----+-----+-----+-----+
|book_id|call_no|keyword|  tfidf|
+-----+-----+-----+-----+
|      1|  C0-0|  赵晓耕| 8.5221|
|      1|  C0-0|   大纲| 7.359|
|      1|  C0-0|   方法| 4.5406|
|      1|  C0-0|   总馆| 0.9375|
|      1|  C0-0|   基本| 0.669|
|      1|  C0-0|   书库| 0.6658|
|      1|  C0-0|   中文| 0.0031|
|      1|  C0-0|   图书| 3.0E-4|
|      2|  I234| 符启林| 9.033|
|      2|  I234|   赵国| 7.6467|
|      2|  I234|   时间| 6.8357|
|      2|  I234|   音乐| 6.0207|
|      2|  I234|   总馆| 0.9375|
|      2|  I234|   基本| 0.669|
|      2|  I234|   书库| 0.6658|
|      2|  I234|   中文| 0.0031|
|      2|  I234|   图书| 3.0E-4|
|      3|B516.41| 塑料|16.1043|
|      3|B516.41| 曙光| 9.033|
|      3|B516.41|   总馆| 0.9375|
```

only showing top 20 rows

TextRank提取关键词：

基于TextRank的关键词提取过程步骤如下：

- 把给定的文本T按照完整句子进行分割，对于每个句子，进行分词（jieba词典）和词性标注处理，并过滤掉停用词，只保留指定词性的单词，如名词、动词、形容词，即，其中是保留后的候选关键词。
- 构建候选关键词图 $G = (V, E)$ ，其中V为节点集，上一步生成的候选关键词组成，然后采用共现关系（co-occurrence）构造任两点之间的边，两个节点之间存在边仅当它们对应的词汇在长度为K的窗口中共现，K表示窗口大小，即最多共现K个单词。根据上面公式，迭代传播各节点的权重，直至收敛。
- 对节点权重进行倒序排序，从而得到最重要的T个单词，作为候选关键词。第二部得到最重要的T个单词，在原始文本中进行标记，若形成相邻词组，则组合成多词关键词。

+-----+-----+-----+-----+-----+-----+					
	book_id	call_no	keyword	textrank	
+-----+-----+-----+-----+-----+-----+					
	1	C0-0	朱红	1.0	
	1	C0-0	总馆	0.959441666295381	
	1	C0-0	方法	0.7333855665564373	
	1	C0-0	书库	0.7177954790548419	
	1	C0-0	图书	0.5351519307210455	
	1	C0-0	中文	0.5327606291736268	
	1	C0-0	社会科学	0.5030445869575579	
	2	I234	总馆	1.0	
	2	I234	孟京辉	0.8441608121096489	
	2	I234	书库	0.8130387273557808	
	2	I234	档案	0.8103230597112169	
	2	I234	中文	0.6506109539398388	
	2	I234	戏剧	0.6179449941811024	
	2	I234	先锋	0.6139648797792653	
	2	I234	图书	0.4564180992142264	
	3	B516.41	叔本华	1.0	
	3	B516.41	总馆	0.7459824864895116	
	3	B516.41	人生哲学	0.7411323928776742	
	3	B516.41	图书	0.5286955076722697	
	3	B516.41	中文	0.5243878372080547	
+-----+-----+-----+-----+-----+-----+					

图书画像结果:

对图书进行计算画像

•步骤:

- 1、加载IDF，保留关键词以及权重计算 (TextRank * IF-IDF)
- 2、合并关键词权重到字典结果
- 3、将tfidf和textrank共现的词作为主题词
- 4、将主题词表和关键词表进行合并，插入表

book_id	call_no	keywords	topics
26	J617.3	Map(图书 -> 1.58437...)	[舞剧, 艺术, 建筑, 阅览室,...]
29	I207.37	Map(图书 -> 2.13892...)	[总馆, 戏曲, 书库, 中文,...]
474	I533.45	Map(馆藏 -> 1.31139...)	[图书, 嘉庚, 挪威, 馆藏,...]
964	D922.504	Map(图书 -> 1.67928...)	[法律, 总馆, 制度, 刘勇,...]
1677	F224.5	Map(图书 -> 1.64540...)	[总馆, 项目计划, 孙军, 书库...]
1697	Q981.1-49	Map(图书 -> 2.10266...)	[孩子, 翔安, 丹尼, 埃利希,...]
1806	I266	Map(图书 -> 1.64186...)	[总馆, 书库, 中文, 图书]
1950	H315.9	Map(图书 -> 1.68068...)	[英汉翻译, 总馆, 陈德彰, 入...]
2040	I247.57	Map(胭脂 -> 4.97531...)	[冯锐, 胭脂, 翔安, 分馆,...]
2214	F279.247	Map(馆藏 -> 1.69534...)	[跨国, 漳州, 企业, 中国,...]
2250	B089.1	Map(王雨辰 -> 9.0329...)	[总馆, 乌托邦, 王雨辰, 书库...]
2453	I247.7	Map(图书 -> 1.62721...)	[翔安, 孙少山, 分馆, 中文,...]
2509	F252	Map(馆藏 -> 1.60010...)	[图书, 张敏, 农产品, 嘉庚,...]
2529	J292.1-09	Map(图书 -> 2.16851...)	[尹旭, 中国, 美学史, 书法,...]
2927	I207.419	Map(图书 -> 1.32265...)	[总馆, 梁羽生, 书库, 中文,...]
3091	Q813.2	Map(图书 -> 2.41375...)	[细胞融合, 图书, 翔安, 分馆...]
3506	TP11	Map(陈根 -> 9.03296...)	[经典案例, 新书, 交互设计,...]
3764	F0	Map(馆藏 -> 1.75318...)	[图书, 嘉庚, 馆藏, 校区,...]
4590	D993.5	Map(图书 -> 2.26724...)	[图书, 船舶, 法学, 分馆,...]
4823	K826.16=76	Map(图书 -> 1.68109...)	[林徽因, 梁思成, 翔安, 费慰...]

only showing top 20 rows

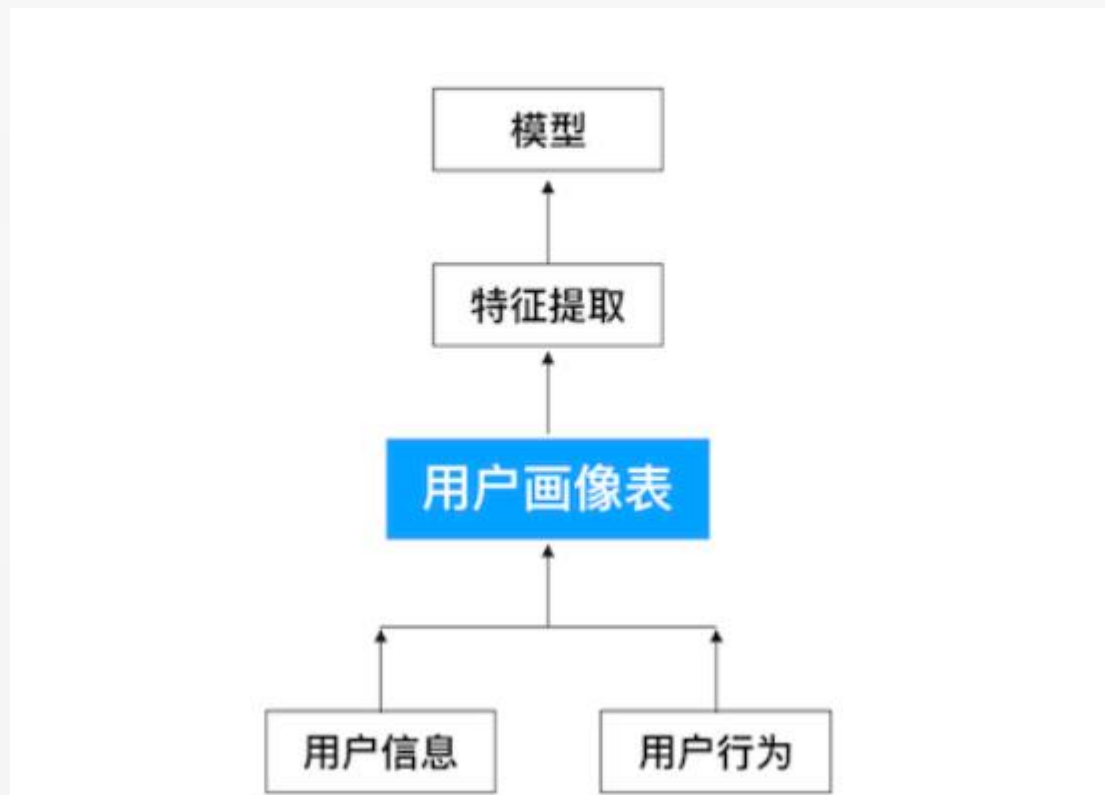
文章相似度word2vec:

实践中word2vec在大量数据下达到的效果更好，离线状态下将图书之间的相似度计算好

- 1、加载某个索书号模型，得到每个词的向量
- 2、获取索书号类别的图书画像，得到图书画像的关键词
- 3、计算得到图书每个词的向量
- 4、计算得到图书的平均词向量即图书的向量
- 5、进行相似度计算

```
+-----+-----+-----+
|book_id|  call_no|          bookvector|
+-----+-----+-----+
|      26|    J617.3| [-0.1411550119519...|
|      29|    I207.37| [-0.0360910377154...|
|     474|    I533.45| [-0.1688659844055...|
|     964|   D922.504| [-0.0193844965542...|
|    1677|    F224.5| [-0.0455365683883...|
|    1697| Q981.1-49| [-0.1101982154262...|
|    1806|     I266| [-0.0455365683883...|
|    1950|    H315.9| [-0.0421454471846...|
|    2040|    I247.57| [-0.1597369369119...|
|    2214|   F279.247| [-0.1354838237166...|
|    2250|    B089.1| [-0.0455365683883...|
|    2453|    I247.7| [-0.1597369369119...|
|    2509|     F252| [-0.1113766044494...|
|    2529| J292.1-09| [-0.0459611634723...|
|    2927|    I207.419| [-0.0311386052053...|
|    3091|    Q813.2| [-0.1466118231415...|
|    3506|     TP11| [-0.0540474951267...|
|    3764|      F0| [-0.1970650230844...|
|    4590|    D993.5| [-0.0547918813037...|
|    4823| K826.16=76| [-0.1132439794891...|
+-----+-----+-----+
only showing top 20 rows
```

用户画像构建流程： reader基本信息表和reader_action行为表



1.对用户行为数据进行处理，合并图书画像中的主题词

book_id	lend_date	ret_date	renew_times	call_no	reader_id	borrowed	call_no	topics
2921	2017-10-26 10:33:53	2017-11-10 15:26:26	1	D923.405	1003	true	D923.405	[邓尧, 律师, 知识产权, 案例...]
10653	2017-02-20 16:21:16	2017-03-02 11:56:32	1	0141.4	1006	true	0141.4	[图书, 嘉庚, 案例, 馆藏, ...]
23710	2017-02-13 18:29:36	2017-04-15 21:08:14	1	B221.5	1025	true	B221.5	[何丽野, 八字, 总馆, 思维, ...]
6059	2017-06-09 08:26:27	2017-07-24 10:22:31	1	K203	1026	true	K203	[单霁翔, 新书, 世纪, 书库, ...]
1366	2017-11-03 16:22:58	2017-11-20 15:54:40	1	K928.6	1030	true	K928.6	[中国, 总馆, 文化, 地名, ...]
7753	2017-10-30 19:34:05	2017-12-01 20:45:25	1	I246.7	1042	true	I246.7	[张爱玲, 翔安, 倾城, 分馆, ...]
21876	2017-03-29 12:33:32	2017-05-28 09:22:00	1	B221.2	1047	true	B221.2	[总馆, 周易, 书库, 中文, 图书]
15879	2017-02-14 20:58:59	2017-04-15 17:05:22	1	I247.58	1061	true	I247.58	[总馆, 笑傲江湖, 书库, 中文...]
6599	2017-11-28 21:47:13	2018-03-04 18:50:14	1	0411.1	1070	true	0411.1	[可视化, 数学, 方程, 总馆, ...]
784	2017-01-02 10:50:05	2017-03-09 10:02:42	1	F279.23	1076	true	F279.23	[模式, 企业, 案例, 新书, ...]
20670	2017-06-27 08:48:26	2017-06-29 08:06:39	1	D922.297	1089	true	D922.297	[资料室, 建筑, 雷明, 中文, ...]
7091	2017-03-20 20:24:17	2017-05-22 17:41:21	1	D923.901	1116	true	D923.901	[杨立新, 总馆, 继承法, 书库...]
20581	2017-06-12 21:19:37	2017-06-13 17:22:26	1	D922.104-44	1126	true	D922.104-44	[行政法, 练习题, 行政诉讼法, ...]
20584	2017-06-12 21:19:42	2017-06-14 10:03:04	1	D925.25	1126	true	D925.25	[疑难, 案例, 陈立, 法学, ...]
10767	2017-10-15 12:41:31	2017-10-22 17:33:47	1	013	1145	true	013	[科技, 数学系, 大学, 新书, ...]
1180	2017-04-14 09:40:45	2017-04-26 15:53:37	1	I234	1156	true	I234	[图书, 高行健, 四重奏, 嘉庚...]
7732	2017-06-18 18:33:26	2017-09-16 18:24:53	1	I247.58	1164	true	I247.58	[图书, 总馆, 书库, 中文, ...]
20531	2017-09-30 11:48:58	2017-12-12 17:16:00	1	D035	122	true	D035	[朱仁显, 总馆, 概论, 公共事...]
9098	2017-05-08 11:05:02	2017-06-08 14:17:56	1	J650.9	1220	true	J650.9	[总馆, 欧洲, 张洪岛, 书库, ...]
24064	2017-09-11 20:12:27	2017-10-14 12:14:58	1	I246.7	1223	true	I246.7	[图书, 张天翼, 总馆, 书库, ...]

2.用户标签权重计算（对每个标签进行打分）：

用户标签权重 = (行为类型权重之和) × 时间衰减

权重参数

```
weightsOfaction = {  
    "read_min": 1,  
    "read_middle": 2,  
    "borrow": 2,  
}
```

计算时间间隔：归还时间-借阅时间

时间衰减: $1/(\log(t)+1)$,t为时间发生时间距离当前时间的大小

根据借阅时间距离和借阅行为分别打分计算权重

召回设计：

一、用户冷启动（解决方案探讨）

1. 非个性化推荐

热门召回：热门图书

新图书召回：新书速递

2. 个性化推荐：

基于内容的协同过滤在线召回：基于用户兴趣画像相似的召回结果，用于个性化推荐

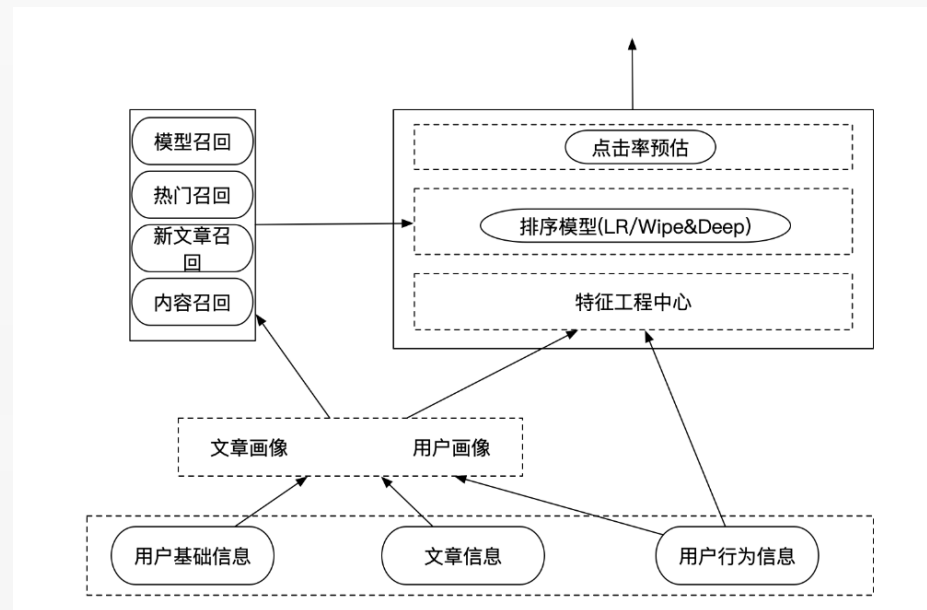
二、后期离线部分（积累一定用户行为数据）

1. 建立用户兴趣画像：包括用户各个维度的兴趣特征

2. 离线部分的召回：

基于模型协同过滤推荐离线召回：ALS

基于内容的离线召回：或者称基于用户画像的召回



1. 基于模型的召回:

目标: 使用ALS模型推荐图书给用户

步骤:

- 1、数据类型转换, borrowed, 用户id和图书id
- 2、ALS模型训练以及推荐
- 3、推荐结果解析处理
进行数据类型反向转换, 根据索引号分类推荐
- 4、推荐结果存储 (Hbase: cb_recall)

```
+-----+-----+-----+
|als_reader_id|      recommendations|reader_id|
+-----+-----+-----+
|          299|[[2009,0.90985286...|      4340|
|          305|[[765,0.90573215]...|       745|
|          496|[[850,0.90344954]...|     1240|
|          558|[[1993,0.90371054...|     4610|
|          596|[[1993,0.9065753]...|     4532|
|          692|[[1993,0.9026662]...|     2430|
|          769|[[854,0.9034734],...|     1031|
|          934|[[854,0.90347224]...|     4383|
|         1051|[[1773,0.90394694...|      317|
```

```
+-----+-----+-----+
|reader_id|   call_no|           book_list|
+-----+-----+-----+
|         1|    D917.6|           [1657]|
|         1|    H319.9|          [21379]|
|         1|    I247.5|[7286, 13965, 11103]|
|         2|    O174.1|           [9172]|
|         3|     TU242|          [14879]|
|         4|    I247.7|          [16638]|
|         5|     TP181|           [5768]|
|         6|    D922.28|         [19645]|
|        10| R195.1-44|         [13377]|
|        11|         G0|         [21187]|
|        12|   H319.4:D|         [3281]|
```

2. 基于内容的召回:

步骤:

1、过滤用户借阅的图书

borrowed=true

2、用户每次操作图书进行相似获取并进行推荐

3. 推荐结果存储 (Hbase: cb_recall)

lend_date	ret_date	renew_times	call_no	book_id	reader_id	borrowed
2017-10-26 10:33:53	2017-11-10 15:26:26	1	D923.405	2921	1003	true
2017-02-20 16:21:16	2017-03-02 11:56:32	1	O141.4	10653	1006	true
2017-02-13 18:29:36	2017-04-15 21:08:14	1	B221.5	23710	1025	true
2017-06-09 08:26:27	2017-07-24 10:22:31	1	K203	6059	1026	true
2017-11-03 16:22:58	2017-11-20 15:54:40	1	K928.6	1366	1030	true
2017-10-30 19:34:05	2017-12-01 20:45:25	1	I246.7	7753	1042	true
2017-03-29 12:33:32	2017-05-28 09:22:00	1	B221.2	21876	1047	true
2017-02-14 20:58:59	2017-04-15 17:05:22	1	I247.58	15879	1061	true
2017-11-28 21:47:13	2018-03-04 18:50:14	1	O411.1	6599	1070	true
2017-01-02 10:50:05	2017-03-09 10:02:42	1	F279.23	784	1076	true
2017-06-27 08:48:26	2017-06-29 08:06:39	1	D922.297	20670	1089	true
2017-03-20 20:24:17	2017-05-22 17:41:21	1	D923.901	7091	1116	true
2017-06-12 21:19:37	2017-06-13 17:22:26	1	D922.104-44	20581	1126	true
2017-06-12 21:19:42	2017-06-14 10:03:04	1	D925.25	20584	1126	true
2017-10-15 12:41:31	2017-10-22 17:33:47	1	O13	10767	1145	true
2017-04-14 09:40:45	2017-04-26 15:53:37	1	I234	1180	1156	true
2017-06-18 18:33:26	2017-09-16 18:24:53	1	I247.58	7732	1164	true
2017-09-30 11:48:58	2017-12-12 17:16:00	1	D035	20531	122	true
2017-05-08 11:05:02	2017-06-08 14:17:56	1	J650.9	9098	1220	true
2017-09-11 20:12:27	2017-10-14 12:14:58	1	I246.7	24064	1223	true

only showing top 20 rows

选取用户ID为1014, 用户的借阅行为和推荐结果如图

信息	结果 1	剖析	状态							
id	reader_no	reader_	book_id	call_no	lend_date	ret_date	renew_times	m_title		
▶ 3834	085CECD5F2FB6D3A48	1014	18766	O6	2017-03-11 15:25:52	2017-03-11 15:26:31	0	合成化学		
10984	085CECD5F2FB6D3A48	1014	19260	F274	2017-04-15 15:56:50	2017-05-20 17:36:19	0	采购管理与库存控制		
26943	085CECD5F2FB6D3A48	1014	20624	D915.204	2017-03-28 17:03:04	2017-04-13 12:37:09	0	修改后的刑事诉讼法实施情况调查		
33597	085CECD5F2FB6D3A48	1014	14131	I247.5	2017-03-04 12:08:27	2017-03-20 09:26:29	0	人面桃花		
38197	085CECD5F2FB6D3A48	1014	6214	F016-44	2017-12-26 19:41:19	2018-01-02 19:14:12	0	微观经济学 (第2版) 习题集		
43571	085CECD5F2FB6D3A48	1014	11809	D99	2017-10-10 14:04:52	2017-11-23 17:53:13	1	国际法原理与案例解析		
44219	085CECD5F2FB6D3A48	1014	23388	D923.404	2017-08-03 10:34:27	2017-10-09 12:15:19	0	知识产权法		
51717	085CECD5F2FB6D3A48	1014	18286	F272.92	2017-12-04 22:01:00	2017-12-18 08:13:50	0	中小企业薪酬体系设计		
61295	085CECD5F2FB6D3A48	1014	15353	I234	2017-10-24 21:17:00	2017-12-26 18:48:19	0	曹禺剧作		

信息	结果 1	剖析	状态							
id	location_name	m_title	call_no	m_call_no	m_isbn	m_author	m_publisher	m_pub_year	doc_type_name	
1132	翔安分馆	叔本华说欲望与幸福	B516.41	B516.41/416.197	978-7-5609-8164-2	(德) 叔本华	华中科技大学出版社	2012	中文图书	
3457	翔安分馆	我读《易经》	B221-49	B221-49/661	978-7-5640-3765-9	傅佩荣	北京理工大学出版社	2010	中文图书	
3964	基本书库-新书区	波兰尼《大转型》与D5		D5/103	978-7-108-04098-5	王绍光	生活·读书·新知三联	2012	中文图书	
4223	翔安分馆	拓扑心理学原理	B84-069	B84-069/187	978-7-301-19602-1	(德) 库尔特·勒	北京大学出版社	2011	中文图书	
4876	基本书库-新书区	公共管理导论	D035	D035/622.211	978-7-300-20740-7	(澳) 欧文·E·休	中国人民大学出版社	2015	中文图书	
7568	总馆基本书库	精英话语与种族歧视	C912.6	C912.6/211.2	978-7-300-13005-7	(荷) 范·戴克	中国人民大学出版社	2011	中文图书	
7710	总馆基本书库	忏悔录	B503.1	B503.1/612	7-100-02282-7, 7-10	(古罗马)奥古斯	商务印书馆	1963	中文图书	
8136	总馆基本书库	资本主义理解史	D091.5	D091.5/016.3/(1), D	978-7-214-05738-9	张一兵	江苏人民出版社	2009	中文图书	
14493	总馆基本书库	新中国民族政策在云	D633.0	D633.0/284	978-7-5161-1655-5	赵新国	中国社会科学出版社	2012	中文图书	
17578	总馆基本书库	马克思主义中国化史	D61	D61/212	978-7-5004-8702-9	梅荣政	中国社会科学出版社	2010	中文图书	
20610	总馆基本书库	楞严经	B942.1	B942.1/278.1	978-7-101-08797-0	赖永海	中华书局	2012	中文图书	
24670	总馆基本书库	西藏生死书	B946.6	B946.6/146.2	978-7-308-08378-2	索甲仁波切	浙江大学出版社	2011	中文图书	

召回和结果

This template is exclusively designed by Fei er creative

使用bootstrap和Springboot构建了一个简易的展示

F 经济

G 文化、科学、教育、体育


H 语言、文字

I 文学

J 艺术


View All

热门借阅



明朝那些事儿

作者: 当年明月编著



挪威的森林

作者: 村上春树著/林少华译



百年孤独

作者: 马尔克斯著 高长荣译



追风筝的人

作者: 卡勒德·胡赛尼著 李继深译



Social
flat design

KNOW MORE

Type

可借

产权性质、企业融资与资源配置效率

作者: 祝继高著

★★★★★

K833.135.38/616.2

我要借阅



Screen
flat design

KNOW MORE

Type

可借


中小企业集群融资新模式论

作者: 高连和著

★★★★★

F276.3/726

我要借阅



Electric
flat design

KNOW MORE

Type

可借

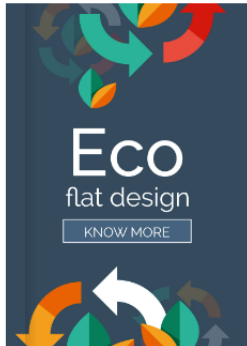
企业融资170种模式及操作案例

作者: 吴维海主编

★★★★★

F279.23/408

我要借阅



Eco
flat design

KNOW MORE

Type

可借

创新型中小企业融资策略

作者: 梁益琳著

★★★★★

F279.243/882.2

我要借阅

类别: Name ▾

数量: 8 ▾

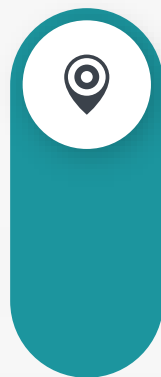
[PART 04]

[总结与展望]



总结与展望

迷茫的探索



不间断的学习



完成项目



总结经验

