2017“达观杯”个性化推荐算法挑战赛

“Sloopy团队” 汇报材料

团队介绍

“sloopy团队(成员：伍可、徐敏、郭志刚、张明静)”，成员主要来自广东电信公司，其中伍可和郭志刚拥有10年以上数据开发和大数据项目管理经验，为“广东省电信大数据专业人才”；多次获得电信集团、省、市大数据竞赛团队和个人奖项。

伍可

主要负责：框架设计、特征提取、算法构建及优化

郭志刚

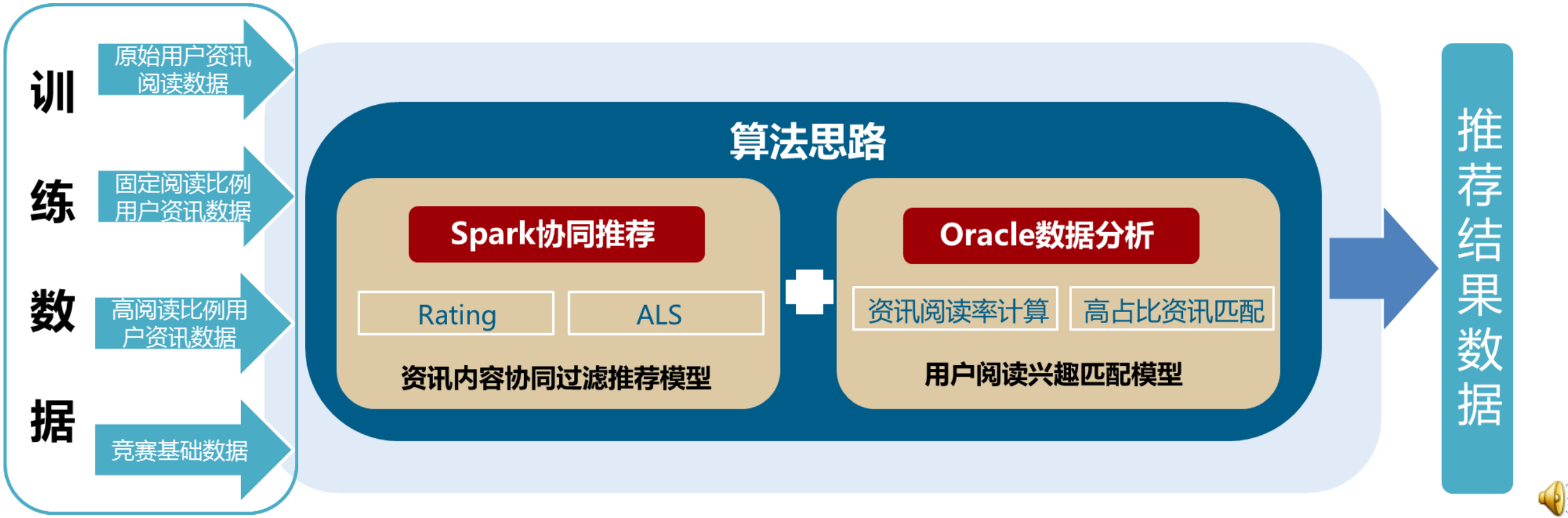
主要负责：数据清洗、特征提取、宽表建立

徐敏

主要负责：数据分析、特征提取

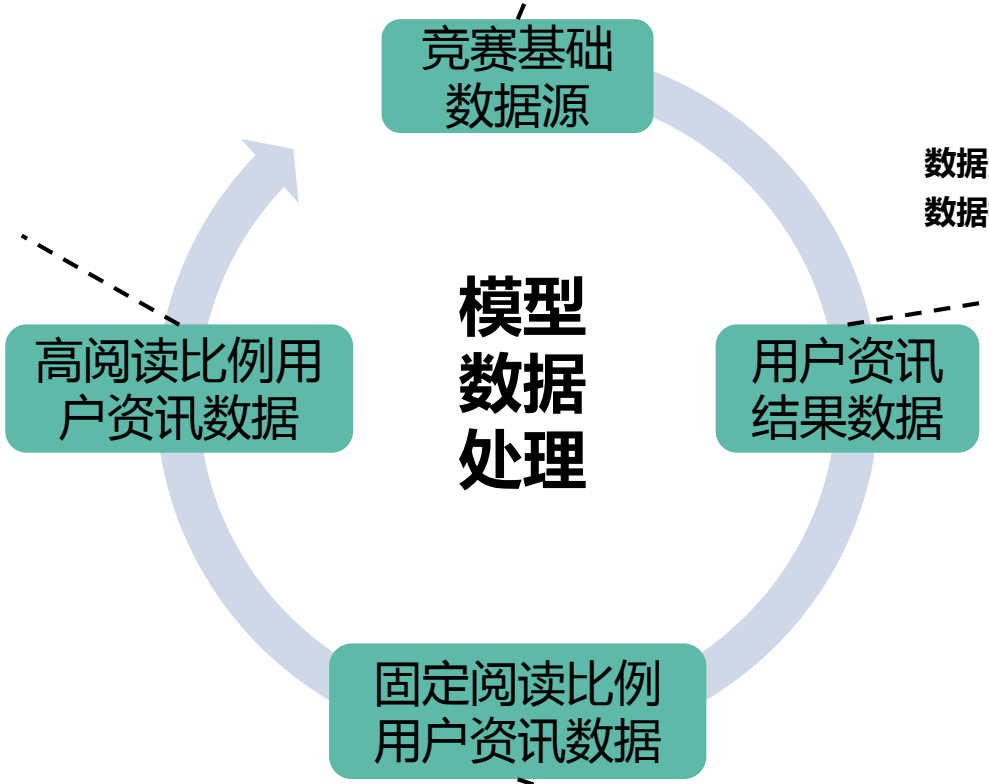
算法思路

数据分析处理 → 模型开发 → 模型优化 → 模型结果



算法思路—数据源分析处理

资讯数据：资讯ID、资讯类别、创建时间；
用户阅读数据：用户ID、资讯ID、资讯类别、用户行为、发生行为时间；
测试数据：测试数据集；
用户数据：待分析用户ID；



数据量：62017
数据格式：用户ID、资讯ID、**资讯评分**
处理规则：计算每个用户对于每个资讯类别阅读比率进行排名，选取用户资讯类别阅读量排名 >=3 && 资讯类别的阅读量 > 3。按照排名选取N+1天资讯类别中最大的资讯ID进行用户和资讯的匹配

数据量：1699479
数据格式：用户ID、资讯ID、**资讯评分**（亮点章节介绍）

数据量：213948
数据格式：用户ID、资讯ID、**资讯评分**
处理规则：每个用户对资讯类别>0.25 && N+1天资讯ID的阅读量 >50 && 用户资讯阅读量 > 1 的用户资讯匹配数据。按照排名选取N+1天资讯类别中最大的资讯ID进行用户和资讯的匹配

算法思路—Spark协同过滤推荐模型2-1

1、模型原理

利用兴趣相投、拥有共同经验群体的喜好来推荐感兴趣的资讯给其他用户；透过合作的机制给予资讯相当程度的评分并记录下来以达到过滤的目的，进而帮助其他用户筛选资讯；本次竞赛中用户评分选取高分优质用户资讯数据。

2、算法说明

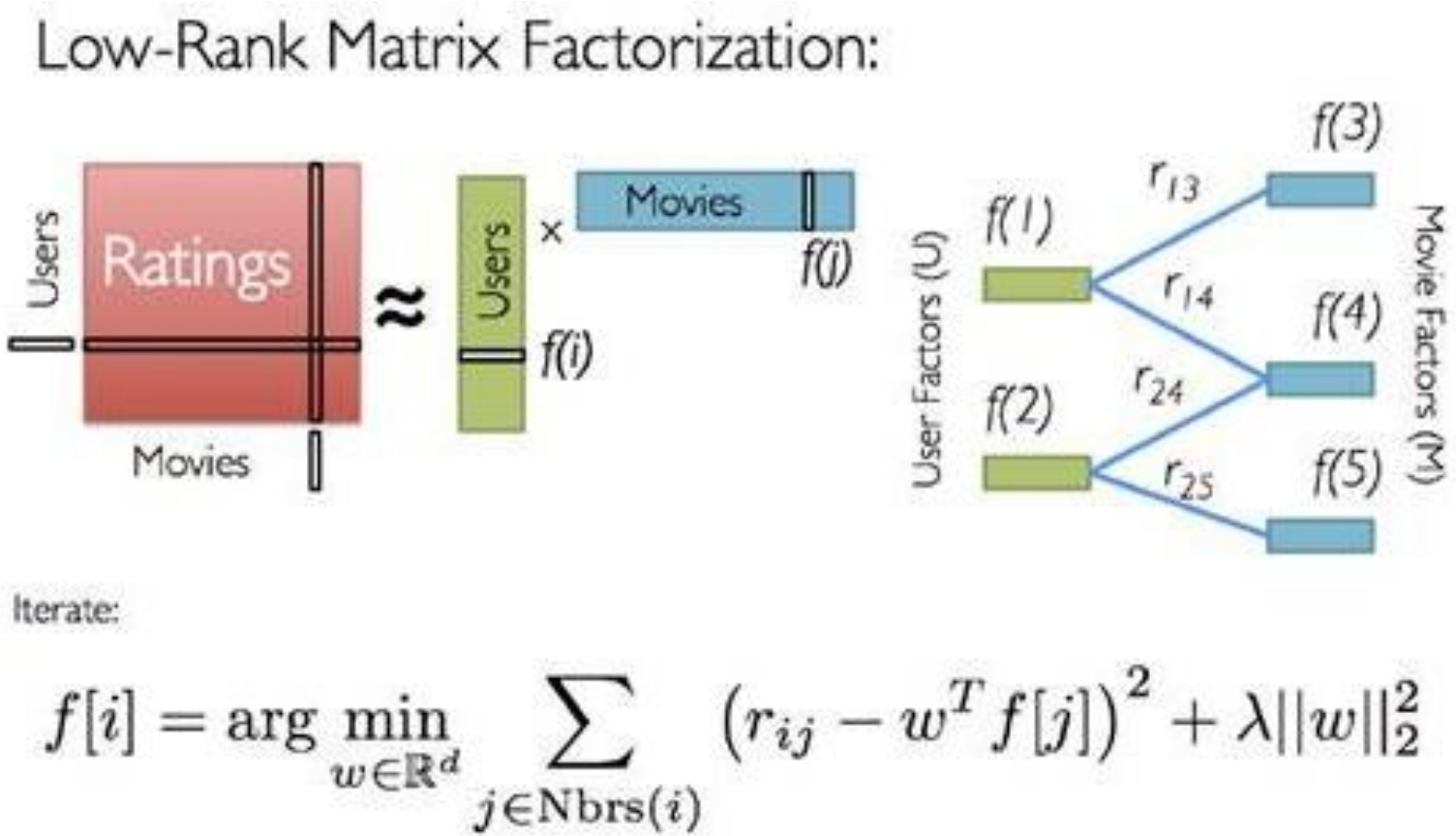
其中用户和资讯通过一小组隐性因子进行表达。Spark-MLLib 使用交替最小二乘法（ALS）来学习这些隐性因子。用户对资讯的偏好，根据应用本身的不同，可能包括用户对资讯的评分、用户阅读资讯的记录等。

3、模型解析

推荐引擎根据不同的推荐机制可能用到数据源中的一部分，然后根据这些数据，分析出一定的规则或者直接对用户对其他资讯的喜好进行预测计算。这样推荐引擎可以在用户进入时给他推荐他可能感兴趣的资讯。

算法思路—协同过滤推荐模型2-2

模型里，用户和资讯被一组可以用来预测缺失项目的潜在因子来描述。特别是我们实现交替最小二乘（ALS）算法来学习这些潜在的因子（下图为互联网取的模型说明）。



算法思路—oracle数据分析处理2-1

1、 N+1天测试数据分析

N+1天资讯阅读占比 = N+1天测试数据的阅读量 * 0.000142889

注： N+1天资讯阅读占比：当天全量用户中有多少比例的用户会阅读该资讯

2、高阅读比例用户资讯数据

根据用户资讯阅读历史，计算用户资讯类别的喜爱程度排名，根据排名在每个资讯类别中匹配资讯阅读占比最高的资讯ID与用户进行数据匹配。

```
Declare
   Cursor cur Is Select user_id From Kesicomp_Candiate;
   inBill Varchar2(1024);

Begin
   for info in cur loop

         inBill := 'Insert Into  tmp_wuk_kesi_empty1 With aa As (Select * From KesiCompt_item_time Where Rowid In(
Select Max(Rowid) From KesiCompt_item_time Where (cate_id,four) In(
Select cate_id,max(four) c_num From KesiCompt_item_time Where cate_id In(Select cate_id From KesiCompt_Train_2_tmp2 Where /*c_num > 20 and*/ user_id = '''
/*And item_id Not In(Select item_id From bs_kesi_result_new_1_1 Where user_id = '''|| to_char(info.user_id)||''')*/ Group By cate_id) Group By cate_id))
Select b.USER_ID, b.C_NUM, b.T_PER, b.T_RANK,a.ITEM_ID, a.CATE_ID, a.FOUR  From aa a,KesiCompt_Train_2_tmp2 b Where a.cate_id = b.cate_id
And b.user_id = '''|| to_char(info.user_id)||''''';

--       Sys.Dbms_Output.put_line(inBill);
          execute immediate inBill;
         commit;
   End Loop;
End;
```

算法思路—oracle数据分析处理2-2

3、最优公式筛选

- 公式1 = 类别阅读量 * N+1天资讯阅读占比
- 公式2 = sqrt（类别阅读量 ）* N+1天资讯阅读占比
- 公式3 = 类别阅读量 * sqrt(sqrt (N+1天资讯阅读占比))* N+1天资讯阅读占比
- 公式4 = 类别阅读量 * sqrt((N+1天资讯阅读占比)* N+1天资讯阅读占比
- 公式5 = = sqrt（类别阅读量 ） * sqrt((N+1天资讯阅读占比)* N+1天资讯阅读占比

根据公式进行最优数据的筛选

4、数据整合校验

- 用户量校验：28501
- 用户资讯匹配校验
- 用户资讯重复记录校验

难点亮点—难点

1、资讯的实时性

资讯在实际的阅读过程中具有实时性和重点信息广泛传播性，同一类型的资讯有可能只在每一个特点的时间范围具有传播性，随着时间的增长，传播会大幅下降；有些资讯具有长时间的传播周期；因此对各类型的资讯要进行分析和类别划分，分析处理过程难度大。

2、资讯的传播爆点

部分资讯并非在出现以后就会进行广泛传播，而是在某一个重要的时间点突然呈现爆发式的传播，这类长时间沉默，突然进行爆炸式传播的资讯较难识别。

3、冷门信息用户阅读而随机性

部分用户存在阅读冷门信息的习惯，其资讯的阅读类别和范围比较随机，没有太多的规律和归一的兴趣爱好，这类用户的资讯匹配难度非常大。

4、阅读量较小用户

用户由于阅读量小，因此会出现部分资讯类别出现很高的阅读占比，在推荐的过程中往往并非用户真实的阅读爱好，由于信息量少，因此资讯推荐难度大。

难点亮点—亮点

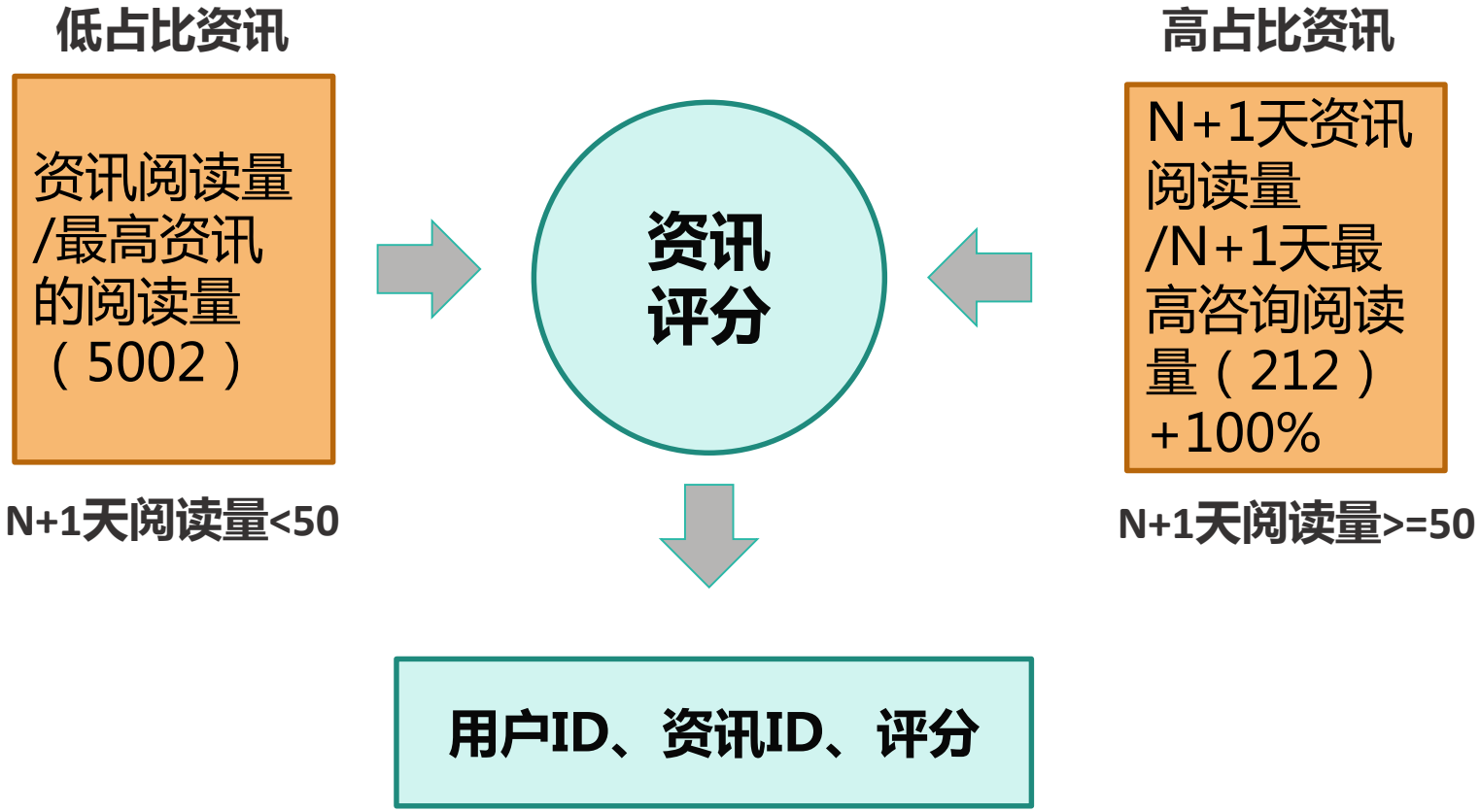
- **N+1天资讯阅读占比系数**

根据测试数据的阅读量计算除**N+1天资讯阅读占比系数=0.000142889**

N+1天测试咨询阅读量	资讯ID	资讯类别	资讯阅读占比	资讯阅读占比系数
212	557579	1_6	0.028178	0.000132915
168	558082	1_11	0.023727	0.000141232
156	558788	1_6	0.021234	0.000136115
134	557167	1_6	0.017655	0.000131754
131	558910	1_1	0.017936	0.000136916
122	556664	1_8	0.018142	0.000148705
111	552472	1_2	0.01673	0.000150721
106	555820	1_6	0.015663	0.000147764
平均资讯阅读占比系数				**0.000142889**

难点亮点—亮点

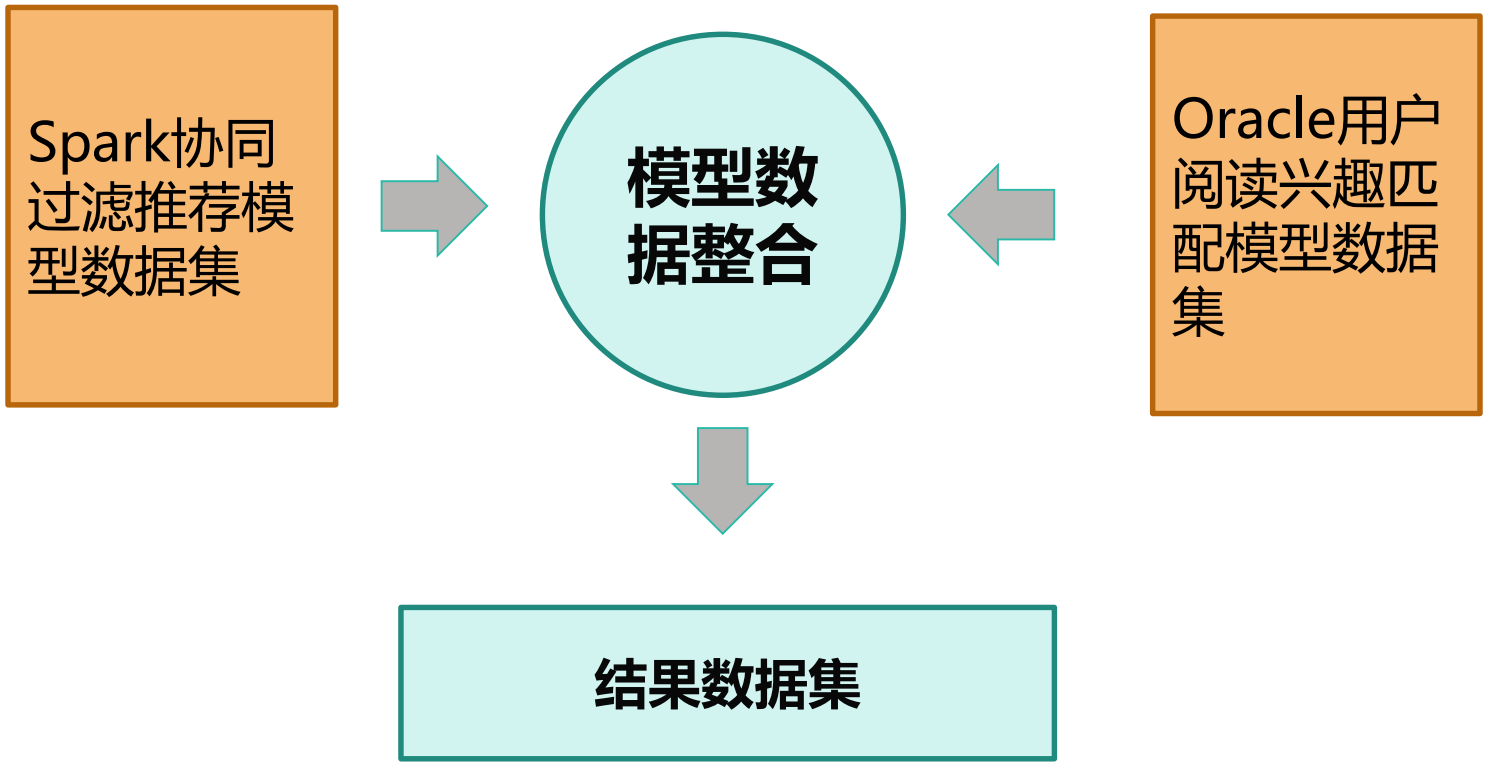
- Spark协同推荐—资讯评分的计算



Spark协同过滤推荐模型需要计算用户对于每个资讯的评分，由于资讯具有实时性和重点资讯的广泛传播性，因此在公式的评分的计算过程中主要采取N+1天测试数据阅读量进行。

难点亮点—亮点

- Spark协同推荐+Oracle数据分析



将Spark协同过滤推荐模型生成的结果数据集与Oracle用户阅读兴趣匹配模型数据集进行数据整合，生成高质量的结果数据；再对遗漏的用户资讯数据使用Oracle部分的数据集进行匹配，生成完整的最终结果数据。

经验总结

竞赛交流

通过竞赛我们的数据分析和挖掘能力得到很大的提升，同时也确立了我们今后努力的方向：从理论、实践、工程融合等方面进一步提升数据挖掘认识和水平；通过与社会其他队伍的交流竞赛，结实大量志同道合的良师益友。

“达观杯”挑战赛

团队协作

团队成员通过无私的合作，利用业余时间完成了挑战赛。竞赛中成员在沟通交流、思路想法、模型处理等多个方面遇到了问题，最终在大家的共同努力下，战胜了所有的困难，竞赛的过程提升了我们团队的协作和项目管理能力，受益终生。

获奖感言

感谢

- 感谢“2017年达观杯个性化推荐算法挑战赛”的主办方：达观数据、Kesci等，为五湖四海的参赛选手提供互相竞技，互相交流、互相学习的机会和平台；
- 感谢“SLOOPY团队”成员在两个月的时间内通力合作、相互理解和齐心协力；
- 感谢家属们的理解和支持，让我们能在业余时间全身心投入到竞赛中；
- 感谢为挑战赛提供保障和支持的幕后英雄；

祝福

- 祝愿达观数据、Kesci等有着创新精神的公司，业务发展能够蒸蒸日上、前景无限；
- 祝愿所有的参赛选手能在事业和学业上更上一层楼，实现自己的人生梦想；

THANKS 感谢聆听