SACC 2014中国系统架构师大会
SYSTEM ARCHITECT CONFERENCE CHINA 2014
发现架构之美

58同城推荐系统设计与实现

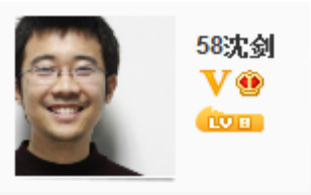
技术中心-沈剑
shenjian@58.com

业务场景-58

- 58目前分类：招聘 => 简历推荐+职位推荐

关于-本技术交流

- 是什么：推荐常用的策略与算法
- 是什么：推荐系统难点+设计+实现
- 关于我：58同城推荐系工程总体设计，推荐系统项目负责人
- @58沈剑



目录

- 推荐简介
- 常用推荐方法
- 系统难点+设计+实现
- 其他

第一章、推荐简介

推荐简介

- **用户**在在某个**场景**下对某个**商品或信息**产生了某种**行为**，**系统**会对**另一些商品或信息**进行推荐
- 要素：

（1）用户 - user

（2）场景 - scene

（3）商品或信息 - item

（4）行为 - action

（5）系统 - recommendation-system

（6）推荐结果集合 - recommendation-result / item-set

系统概貌

- **用户**在在某个**场景**下对某个**商品或信息**产生了某种**行为**，**系统**会对**另一些商品或信息**进行推荐
- 举例：用户在58同城发布了一份简历

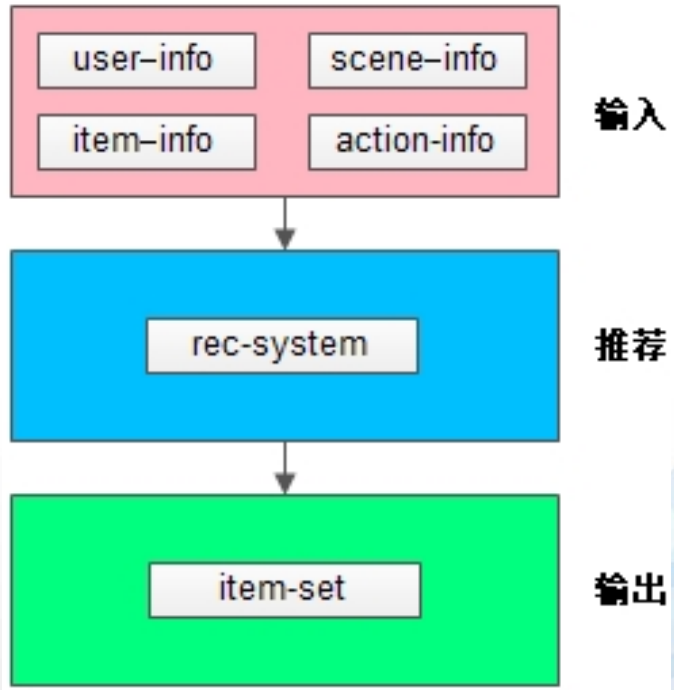
user-info：uid

scene-info：entry、local、cateid

item-info：jid

action：post

item-set：set<zid>



推荐与推荐系统

第二章、常用推荐方法

协同过滤-CF

- 协同过滤：collaborative filtering Recommendation
- 原理：用户的相似喜好进行推荐
- 举例：商家下载简历的推荐

	jid1	jid2	jid3	jid4	jid5	jid6	…	jid1000w
uid1	yes	yes	yes				…	
uid2	yes	yes	yes	yes			…	
uid3	yes	yes	yes		yes	yes	…	
uid4				yes		yes	…	
…								
uid100w								

内容推荐

- 内容推荐：content-based Recommendation
- 原理：抽取共有属性
- 举例：商家下载简历的推荐
- 步骤：

（1）历史行为收集

（2）id详情查询

（3）共性内容挖掘（行为+场景）

（4）推荐

历史行为
jid1，download
jid2，download
zid1，post

详情
jid1（司机，北京，8000月薪，5年经验，NULL）
jid2（司机，北京，NULL，2年经验，硕士研究生）
zid1（司机，北京，NULL，3年经验，NULL）

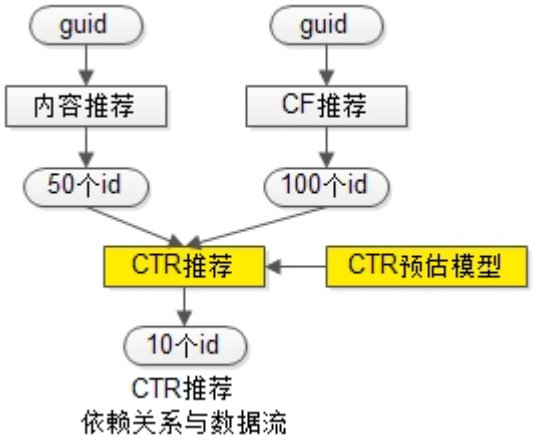
共性
（司机，北京，NULL，2年经验，NULL）

综合排序-CTR预估

- CF算法推荐了50个jid，内容推荐算法推荐了100个jid，最终页面只需要返回5个jid，如何返回？哪个排前面？
- 综合排序
- 什么决定综合排序？
- CTR由什么决定？
- CTR预估打分公式？

用户产品：ctr
CPC商业产品：ctr * price
CPA商业产品：ctr

地区+职位+薪酬范围+工作经验+学历
发帖时间+是否下载过+是否浏览过+...



ctr-score = a*f(地区) + b*f(职位) + c*f(薪酬) + d*f(工作经验) + e*f(学历)
+f*f(发帖时间) + g*f(是否下载过) + h*f(是否浏览过) + ...

第三章、推荐系统难点+设计+实现

线下+线上的系统

- 线下系统
- 线上系统
- 几个问题

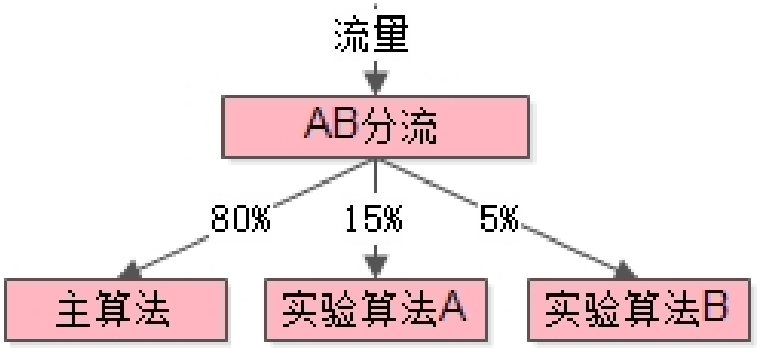
（1）线下数据如何存储？

（2）线上数据如何存储？

（3）如何进行数据转化？

支持实验的系统

- 如何做算法测试？
- 如何快速支持一个算法平台？
- 如何实现分流AB测？



支持实验的系统

支持实验的系统-分流平台

- 作用？需求？

（1）支持random分流（2）支持取模分流（3）支持与或非表达式

（4）支持集合操作（5）配置热加载

- 支持哪些属性的与或非，集合操作？

```
# 招聘数据路径
zhaopin.datapath: /opt/data/zhaopin/
# 招聘组合筛选条件
# 以|分隔，A表示实验平台，1表示优先级，第三部分表示组合筛选条件，多个条件以逗号分隔
zhaopin.default:A|70
zhaopin.condition: A|1|entry.entryDeviceType=IOS,visitor.visitorUserid^uid_whitelist
zhaopin.condition : A|2|visitor.visitorCookieid=58tongcheng
zhaopin.condition: B|3|visitor.visitorUserIP=127.0.0.1
zhaopin.condition: A|4|business.businessEntityId%5%0
zhaopin.condition: B|5|default
```

支持实验的系统-推荐内核

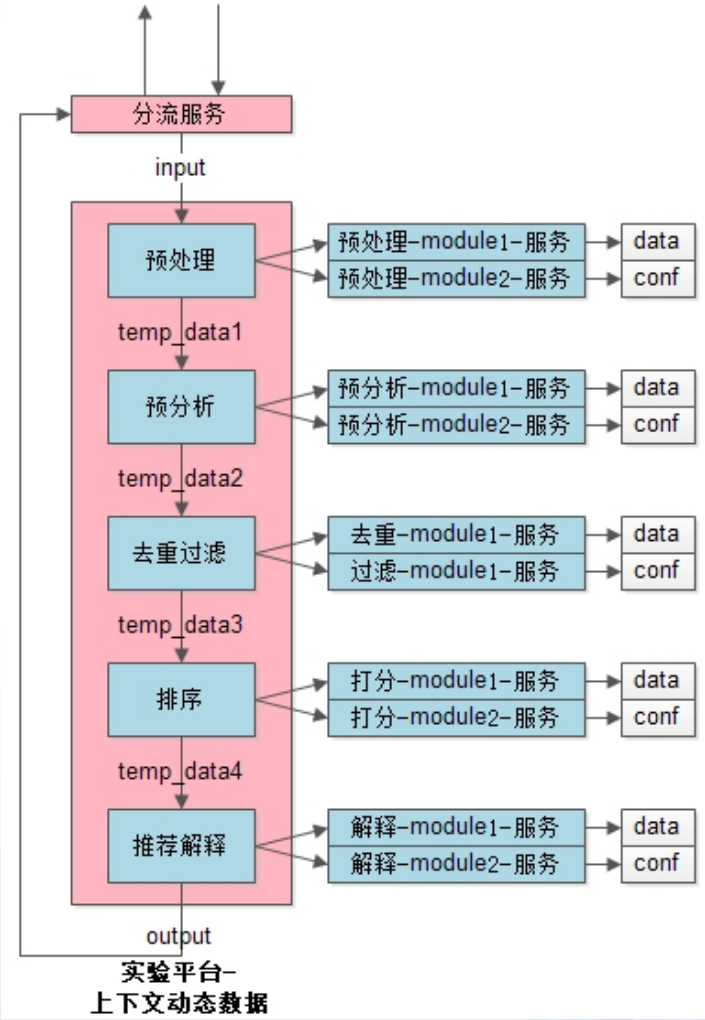
- 作用？需求？
- 设计与实现

（1）算法平台的抽象

（2）实验平台的扩展

（3）上下文动态数据扩展

（4）异步框架与状态机



实时计算的检索系统

- 作用？
- 通用需求

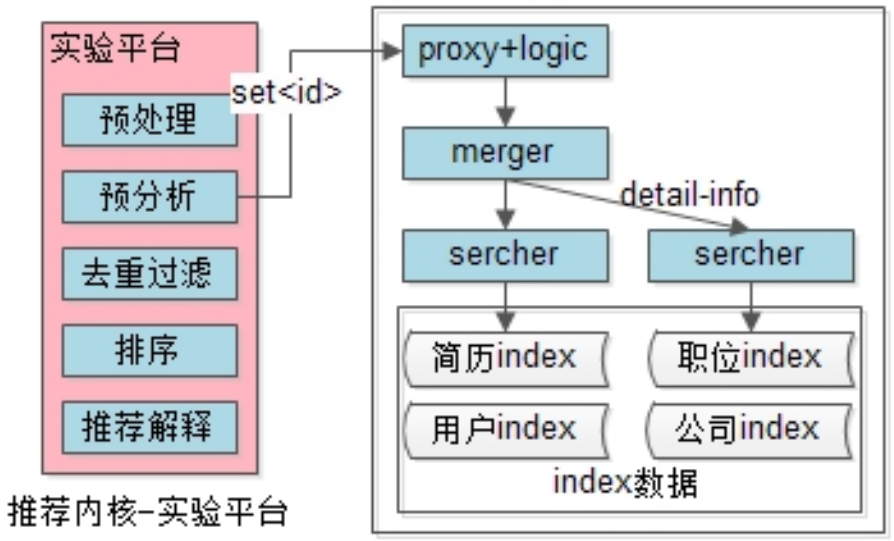
（1）正排数据的存储

（2）倒排索引的存储

（3）数据的更新

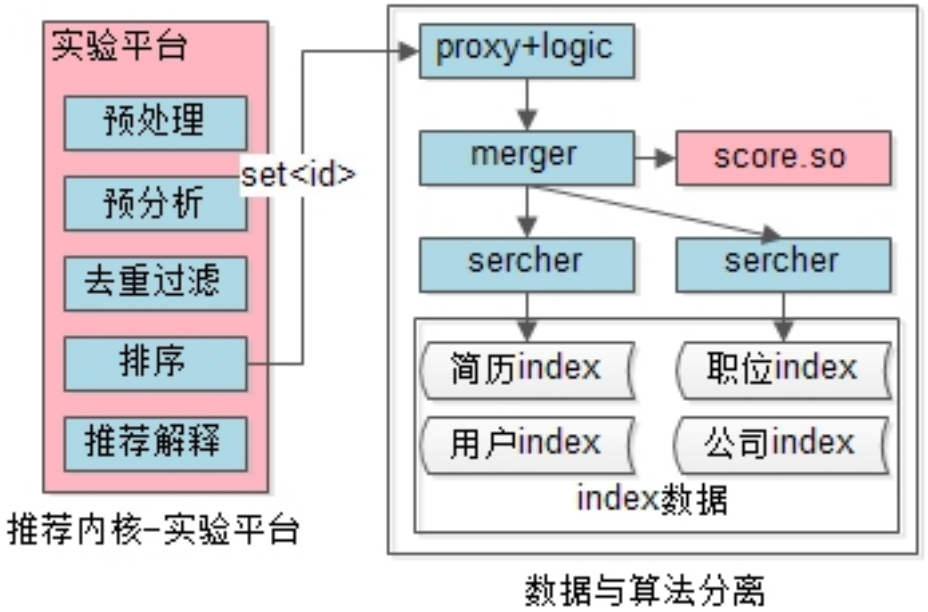
（4）map-reduce的信息查询

- 业务需求



工程+算法的系统

- 如何让工程和算法解耦？
- 线下算法如何分离？
- 实验平台算法如何分离？
- 排序打分算法如何分离？



效果实时监测的系统

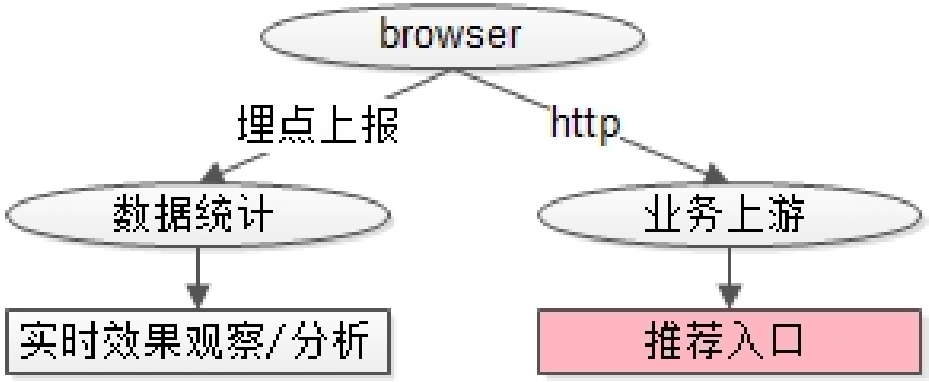
- 作用？
- 步骤

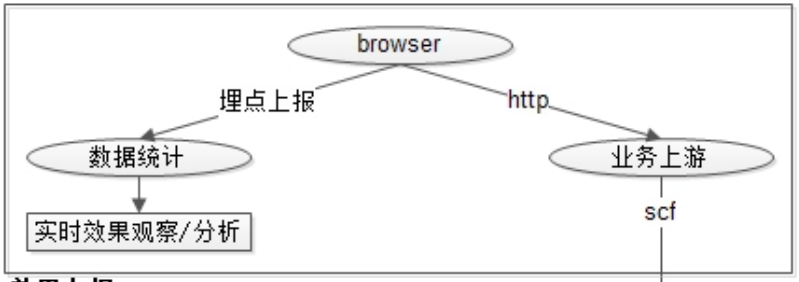
（1）结果展示

（2）结果上报，点击上报

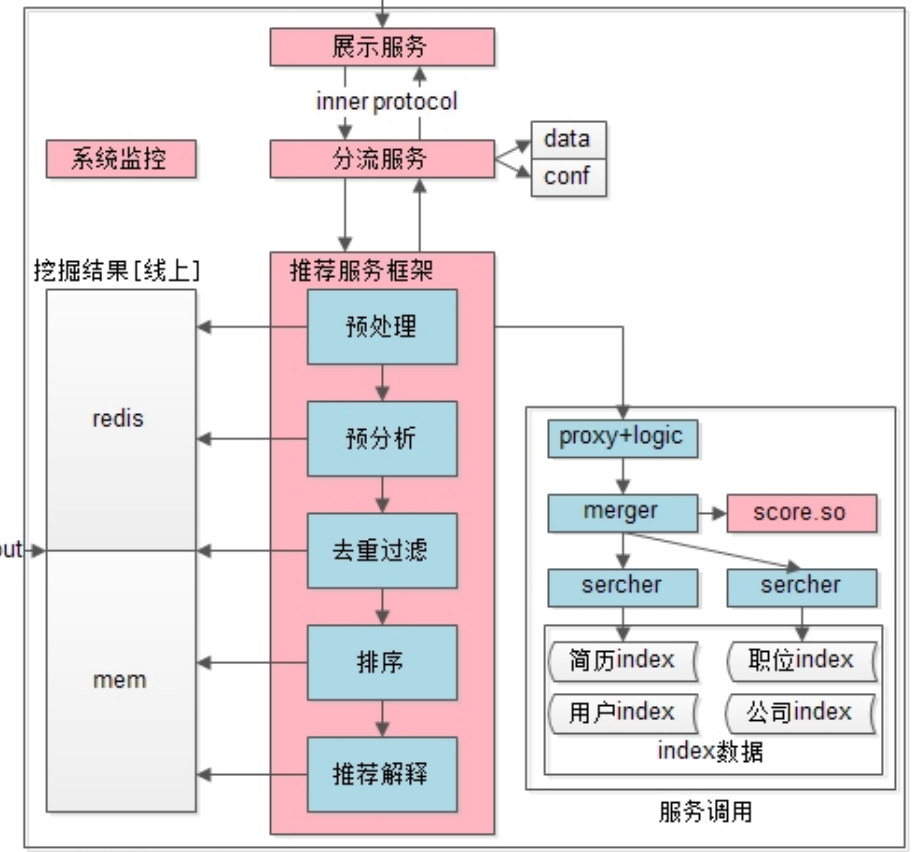
（3）数据收集，数据统计

（4）数据展示



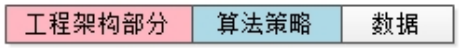


效果上报



推荐系统-线上部分

推荐系统-线下部分



推荐系统-图示

58同城推荐系统设计与实现

- 谢谢大家！
- @58沈剑