系统流程说明：

1、kettle数据抽取，更新ETL\_TEMP表时间字段

2、tag服务每20分钟执行调度任务去获取抽取的最新时间，再根据时间（抽取结束1-2小时内）和redis中的处理标记DEAL\_FLAG为POST\_OVER\_FLG进行打标签

将用户和知识标签存入IRE\_USER\_FOLLOW和IRE\_KNOWLEDGE\_INFO表中

主要处理如下数据

知识数据：筛选表CALTKS\_META\_KNOWLEDGE中ktypeid为6与2的（即质量问题案例与标准规范），分别从关联表中，质量案例表CALTKS\_DK\_ZHILIANGWENTIANLI,相关字段使用textrank抽取关键词，从IRE\_TAG\_WORD中过滤word字段，在该词表中的则为标签，放入到IRE\_KNOWLEDGE\_INFO表 tag\_keyword字段中

过滤配置



IRE\_KNOWLEDGE\_INFO表字段对应来源

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 字段名 | 类型 | 注释 | 源表 | 源字段名 |
| ID | VARCHAR2 | id | CALTKS\_META\_KNOWLEDGE | ID |
| TITLE | VARCHAR2 | 标题 | CALTKS\_META\_KNOWLEDGE | TITLE\_NAME |
| CREATE\_TIME | TIMESTAMP | 时间 | CALTKS\_META\_KNOWLEDGE | UPLOAD\_TIME |
| AUTHOR | VARCHAR2 | 作者 | 无 | 无 |
| SECURITY\_LEVEL | VARCHAR2 | 密级 | CALTKS\_META\_KNOWLEDGE | SECURITY\_LEVEL |
| ABSTRACT\_TEXT | VARCHAR2 | 摘要 | CALTKS\_META\_KNOWLEDGE | ABSTRACT\_TEXT |
| ORG | VARCHAR2 | 组织 | 无 |  |
| KTYPE | VARCHAR2 | 知识类型 | CALTKS\_META\_KNOWLEDGE | KTYPEID |
| TAG\_MODEL | VARCHAR2 | 相关型号 | 暂时未用 |  |
| KNOWLEDGE\_SOURCE | VARCHAR2 | 来源 | CALTKS\_META\_KNOWLEDGE | ID |
| TAG\_KEYWORDS | VARCHAR2 | 关键词标签 | 筛选表CALTKS\_META\_KNOWLEDGE中ktypeid为6与2的（即质量问题案例与标准规范），分别从关联表中，质量案例表CALTKS\_DK\_ZHILIANGWENTIANLI,相关字段使用textrank抽取关键词，从IRE\_TAG\_WORD中过滤word字段，在该词表中的则为标签，放入到IRE\_KNOWLEDGE\_INFO表 tag\_keyword字段中 |  |
| TAG\_DEVICE | VARCHAR2 | 相关设备 | 暂未使用 |  |
| SOURCE\_ID | VARCHAR2 | 源表id | CALTKS\_META\_KNOWLEDGE |  |
| URL | VARCHAR2 | url地址 |  |  |
| DOMAIN | VARCHAR2 | 领域 |  |  |
| LAST\_TIME | TIMESTAMP | 更新时间 |  |  |

IRE\_USER\_FOLLOW表，先从ADMIN\_USER表中过滤ORG\_CODE部门

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 字段名 | 类型 | 注释 | 源表 | 源字段名 |  |
| ID\_NUM | VARCHAR2 | 用户证件号 | ADMIN\_USER | P\_ID |  |
| USER\_NAME | VARCHAR2 | 姓名 | ADMIN\_USER | NAME |  |
| POST | VARCHAR2 | 岗位 | 无 |  |  |
| USER\_JOB | VARCHAR2 | 职务 | PERSON\_POST | VDEFL |  |
| FOLLOW\_MODEL | VARCHAR2 | 最近关注的型号 | 暂未使用，放到FOLLOW\_PRO下 |  |  |
| FOLLOW\_DEVICE | VARCHAR2 | 最近关注的设备分系统 | 暂未使用，放到FOLLOW\_PRO下 |  |  |
| FOLLOW\_PRO | VARCHAR2 | 关注的专业关键词 | 根据ADMIN\_USER的id从ZZ\_MESSAGE\_INFO表中筛选sender和FILETYPE=3的数据，从中获取msg\_info字段，使用textrank抽取关键词 |  |  |
| SECRET\_LEVEL | VARCHAR2 | 密级 | ADMIN\_USER | SECRET\_LEVEL |  |
| ORG\_CODE | VARCHAR2 | 组织 | ADMIN\_USER | ORG\_CODE |  |
| ORG\_NAME | VARCHAR2 |  | ADMIN\_USER | ORG\_NAME |  |
| LAST\_TIME | TIMESTAMP | 更新时间 |  |  |  |

3、calc计算服务每20分钟执行调度任务，根据redis中的处理标记为TAG\_OVER\_FLG按规则进行计算产生匹配结果放入到redis中

根据org\_code筛选去取ire\_user\_follow表用户，与ire\_knowledge\_info进行匹配计算

过滤配置



计算规则如下

1. 兴趣标签直接匹配：对于每个人，获取所有的兴趣标签，当兴趣标签在3个以上时，利用兴趣标签进行匹配，从文档中找到命中相关标签的文档，命中一个标签的相似度加0.3。（当标签数>=1加到推荐结果中）
2. 语义相似度模糊计算：当直接匹配结果小于5个，利用”岗位信息+个人兴趣标签（ire\_user\_follow表字段user\_job+tag\_keywords ）“与物品标题、物品关键词（ire\_knowledge\_info表字段title+tag\_pro）做相似度计算，得出相似度值大于0.3的结果。利用语义相似度计算引擎计算结果。（使用tf-idf进行字段相似度计算，在上述标签命中的结果集中增加语义相似度命中的结果）
3. 部门信息匹配：当以上匹配结果仍小于5个，利用作者标签，选取10个左右的相关科室的物品信息。（缺相关数据，暂未加入）

得到符合规则的匹配数据存入redis中

4、post根据定时调度时间在处理标记为CALC\_OVER\_FLG执行推送，推送时间按照quartz配置进行更改

根据org\_code筛选去取ire\_user\_follow表用户，从redis中获取对应用户的匹配结果数据前5条，若有匹配结果则进行推送



推送成功的数据将放入IRE\_RECOMM\_LOG表中记录用户和对应knowledge的id，下次不再重复计算匹配和推送

5、相关接口为手动执行产生每个过程的数据到对应表中，不去做处理标记判断

Post模块

irecommpost/getresults？size=10&current=1&id\_num=11 接口根据用户身份证先从ire\_user\_follow进行模糊匹配再根据证件号从redis获取计算匹配的最终推送结果



结果如下

Rsource即用户与知识命中的标签

Bz为命中的来源方式（标签和相似度）；

relevancy为分数

