# 顶会论文检索和热词分析工具

# 需求分析与概要设计

## 项目说明

## 项目目标：

本项目是一个Web应用，帮助学者快速查找ICCV, CVPR, MACV三大顶级会议的的论文，展示顶会热词，根据用户偏好做推荐。

## 软硬件环境需求

软件的运行环境规定如下：

操作系统：服务器端可以运行在Windows7以上的操作系统

数据库：Elasticsearch 8, MySql 8以上

网络：客户端需要联网操作，可以在支持css3,html5的高级浏览器上正常运行.

## 使用的关键技术：

（1）使用Elsticsearch做论文存储和全文检索

（2）使用Html Agility Pack和Quarz实现爬虫和任务调度

（3）使用python实现基于热词和用户历史记录的推荐系统

（4）使用apache echarts实现数据可视化

## 需求分析

## 系统用例

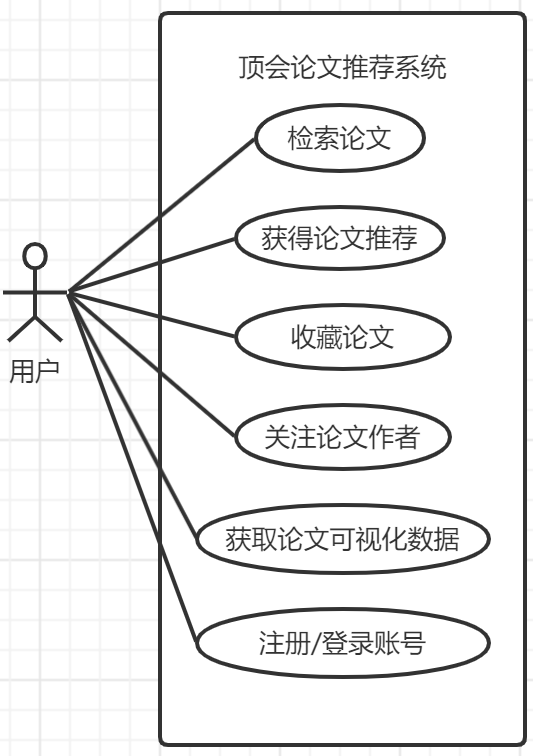


图 系统用例图

1. **注册/登录账号**

**参与者**：用户

**基本事件流**：首次登录需填写部分必要信息，例如近期研究方向、有兴趣了解的领域。顶会论文作者注册时可以进行认证。

1. **浏览、检索、收藏论文**

**参与者**：用户

**基本事件流**：用户在网站主界面可以依据主题检索、浏览可能需要的论文。论文可以被收藏至指定的收藏夹中。

1. **关注论文作者**

**参与者**：用户

**基本事件流**：对于感兴趣的论文作者，用户可以关注并获取其发表论文的动态

1. **获得推荐论文**

**参与者**：用户

**基本事件流**：用户进入网站推荐界面，可以获得推荐阅读的论文。这些论文来自于冷启动论文列表，以及根据用户信息（用户收藏、检索）所计算的论文。

1. **获得可视化数据**

**参与者**：用户

**基本事件流**：用户可在数据统计界面浏览可视化数据，包括以列表、统计图、词云（、知识图谱）等方式呈现的数据。

## 业务流程

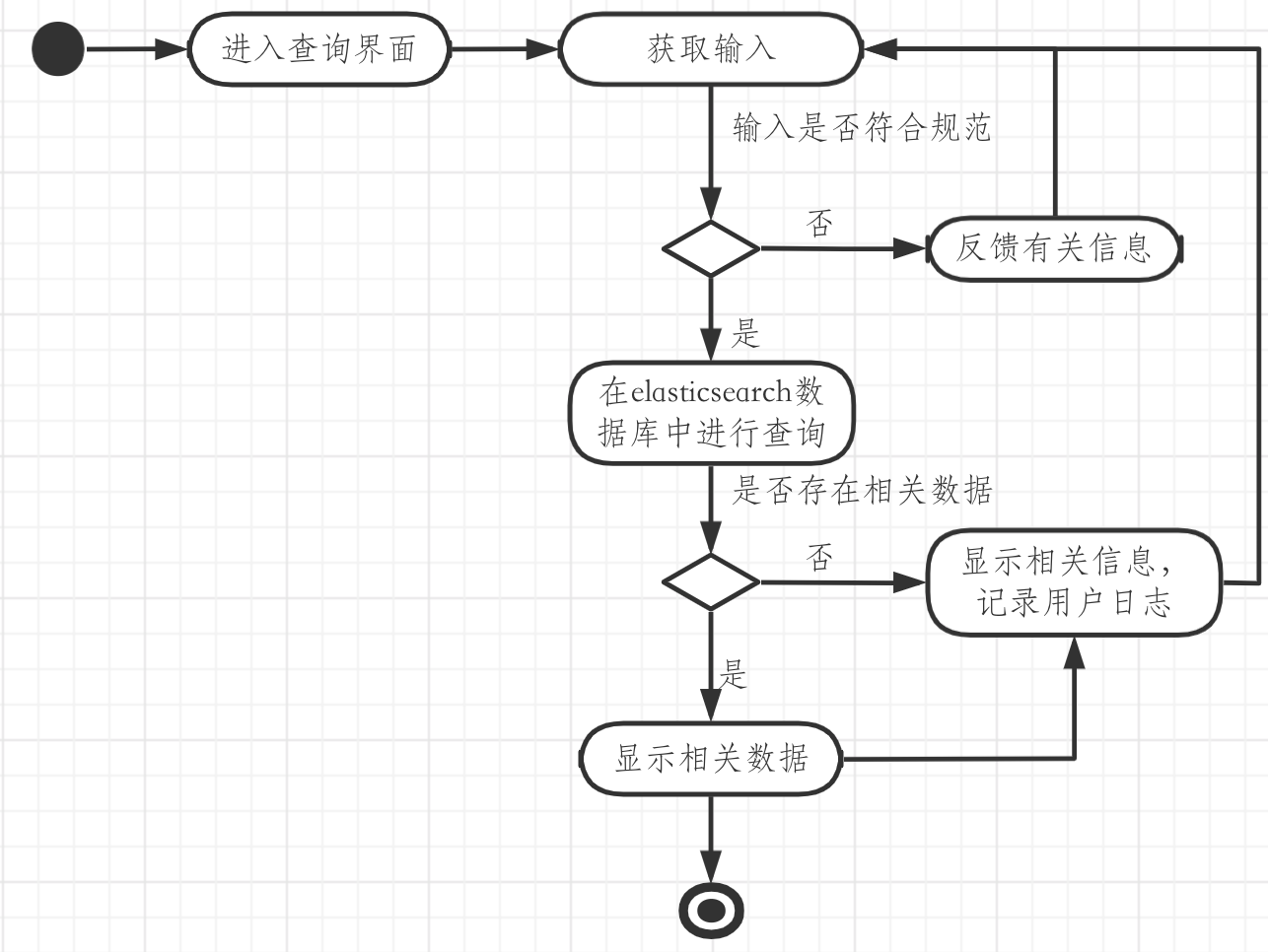
****

图 查询流程图

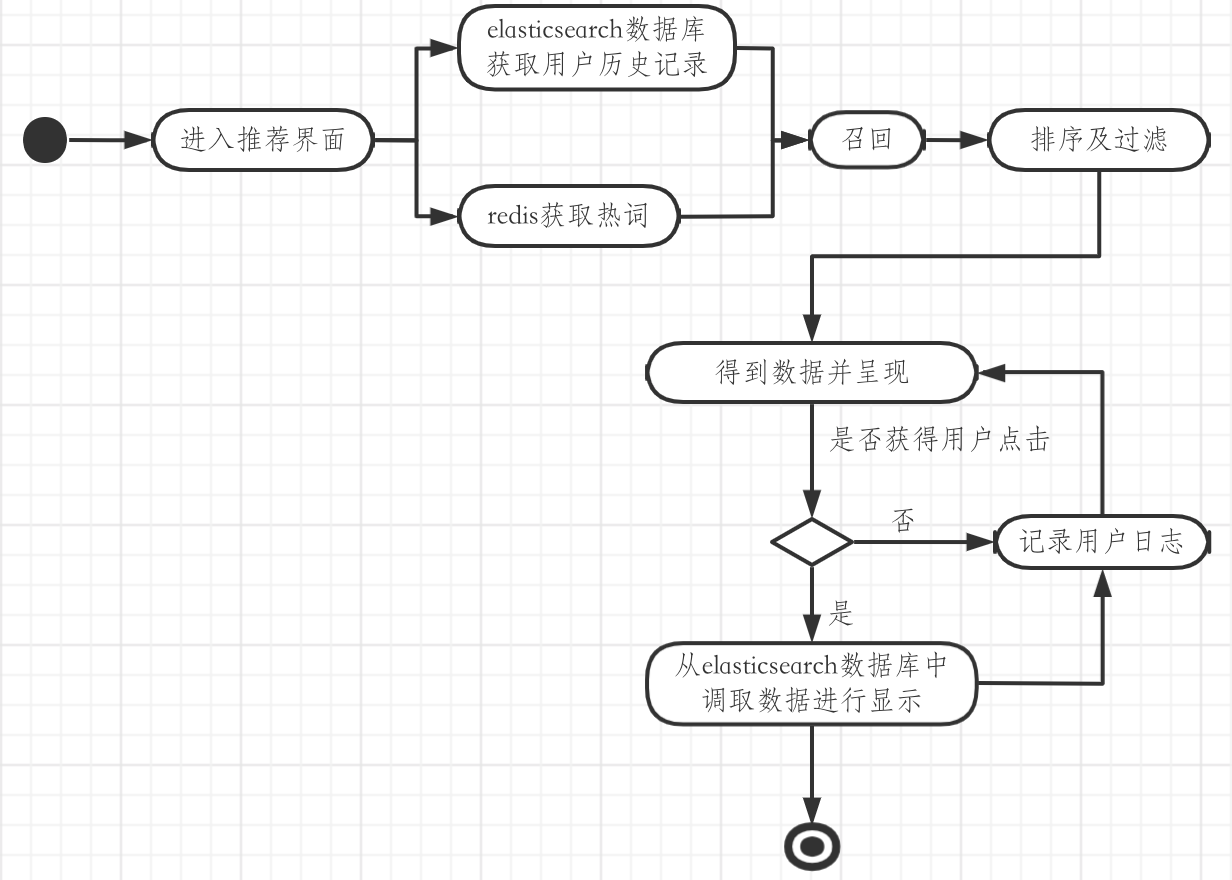
****

图 推荐流程图

## 概要设计

## 功能模块设计

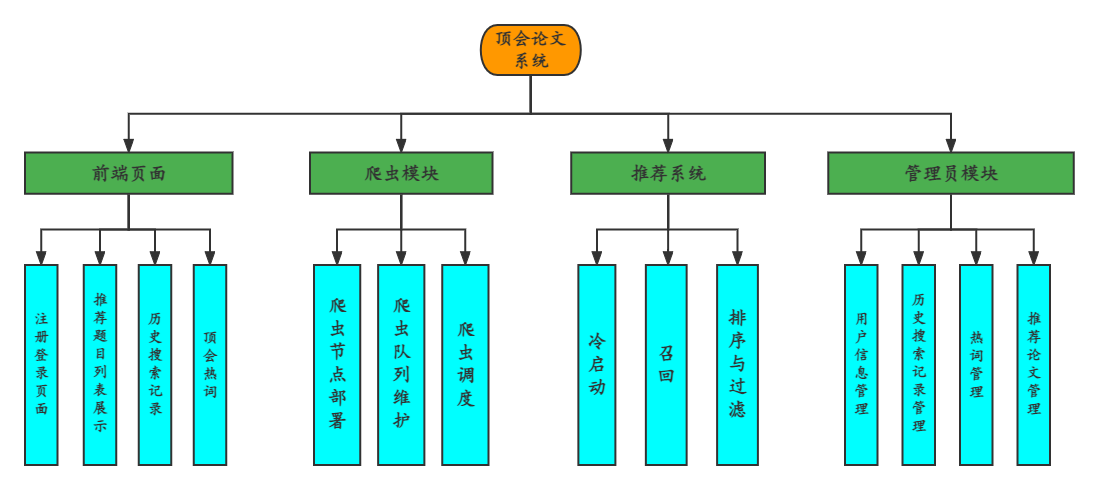


图 功能模块图

1. **注册登录**

**输入：用户名、密码**

**输出：注册(登录)成功(失败)页面**

**功能概述：** 用户输入用户名和密码进行注册(登录)，如果成功返回注册(登录)成功的界面，否则返回注册(登录)失败的界面

1. **查看论文**

**输入：用户发送查看推荐论文的请求**

**输出：推荐的论文题目列表展示界面**

**功能概述：**用户发送查看推荐论文的请求，前端将请求发送给后端处理，后端接受请求调用推荐系统模块进行推荐，将推荐结果发送给前端，前端将结果渲染在Web界面上。

1. **历史搜索记录**

**输入：用户发送查看历史搜索记录的请求**

**输出：用户历史搜索记录展示界面**

**功能概述：** 用户发送查看历史搜索记录的请求，前端将请求发送给后端处理，后端接受请求后调用管理员模块中历史搜索记录管理，查询用户历史搜索记录，将查询结果发送给前端，前端将结果渲染在Web界面上。

1. **顶会热词**

**输入：用户发送查看顶会热词的请求**

**输出：顶会热词展示界面**

**功能概述：**用户发送查看顶会热词的请求，前端将请求发送给后端处理，后端接受请求后调用管理员模块中定会热刺管理，根据用户要求查询对应热词，将查询结果发送给前端，前端将结果渲染在Web界面上。

1. **爬虫节点部署**

**输入：**

**输出：**

**功能概述：**

1. **爬虫队列维护**

**输入：**

**输出：**

**功能概述：**

1. **爬虫调度**

**输入：**

**输出：**

**功能概述：**

1. **冷启动**

**输入：**

**输出：**

**功能概述：**解决系统冷启动问题，包括物品冷启动和用户冷启动。对于物品冷启动，基于协同过滤更新物品相似度表。对于用户冷启动，先提供非个性化的推荐，搜集到一定量的用户行为数据之后，再转为个性化推荐。

1. **召回**

**输入：候选论文集**

**输出：召回后的论文集**

**功能概述：** 输入候选论文集，根据用户的兴趣和历史行为，筛选出找回后的论文集。候选集规模大、模型和特征简单、速度快，尽量保证用户感兴趣数据多召回。

1. **排序与过滤**

**输入：召回后的论文集**

**输出：用户推荐的论文集**

**功能概述：**输入召回后的论文集，通过排序模型给论文打分，返回TOPN的论文集用作推荐。

1. **用户信息管理**

**输入：**

**输出：**

**功能概述：** 对用户信息的增删改查以及鉴权。

1. **历史搜索记录管理**

**输入：**

**输出：**

**功能概述：**对用户历史搜索记录的增删改查。

1. **热词管理**

**输入：**

**输出：**

**功能概述：** 对论文热词的增删改查。

1. **推荐论文管理**

**输入：**

**输出：**

**功能概述：**对推荐系统产生的推荐结果进行查询。

## 核心类图

1 Elasticsearch索引设计

Paper库： 用于存放论文，主要字段有：id, 标题，作者，摘要，发布时间，原文链接，pdf链接，会议名，补充资料链接，被引用数，收藏数，阅读数，热度值，在https://openaccess.thecvf.com/上爬取

Github: 用于存放论文代码，主要字段有：id，论文标题，Github标题，Github链接，在https://github.com/和https://arxiv.org/上爬取

RelatedPaper: 用于存放相关论文，主要字段有id, 论文标题，相关论文标题，相关论文链接，在https://www.connectedpapers.com/上爬取

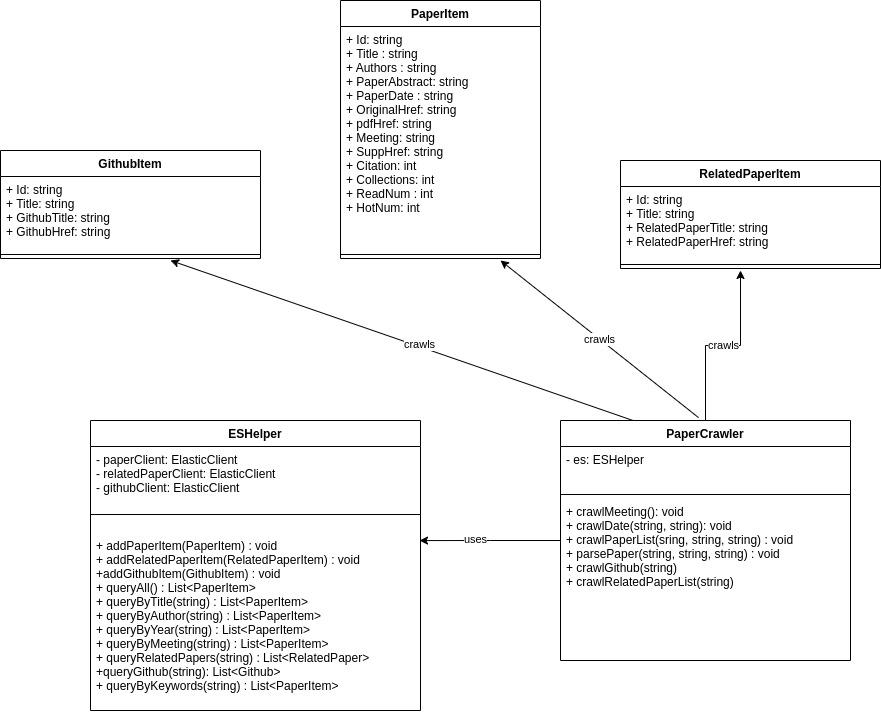


图 爬虫和Elasticsearch数据库

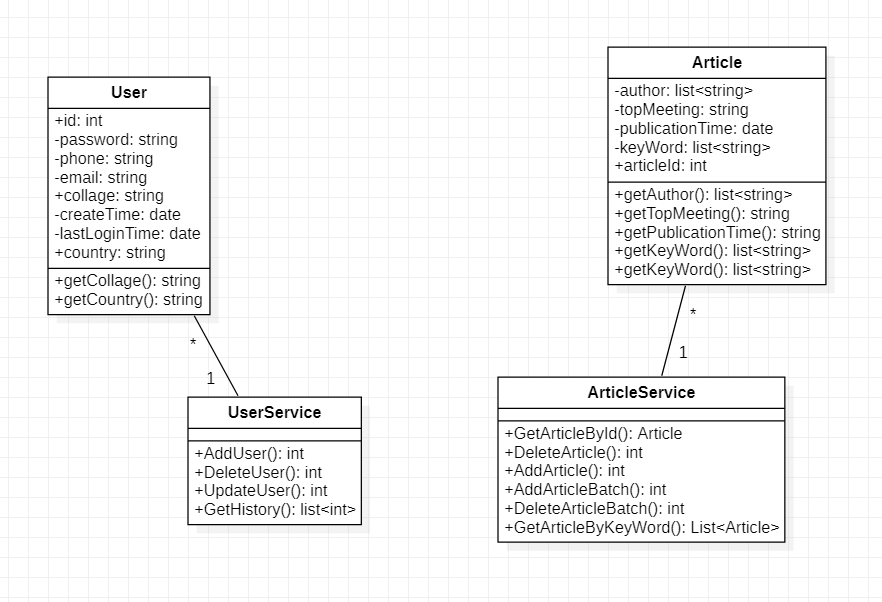
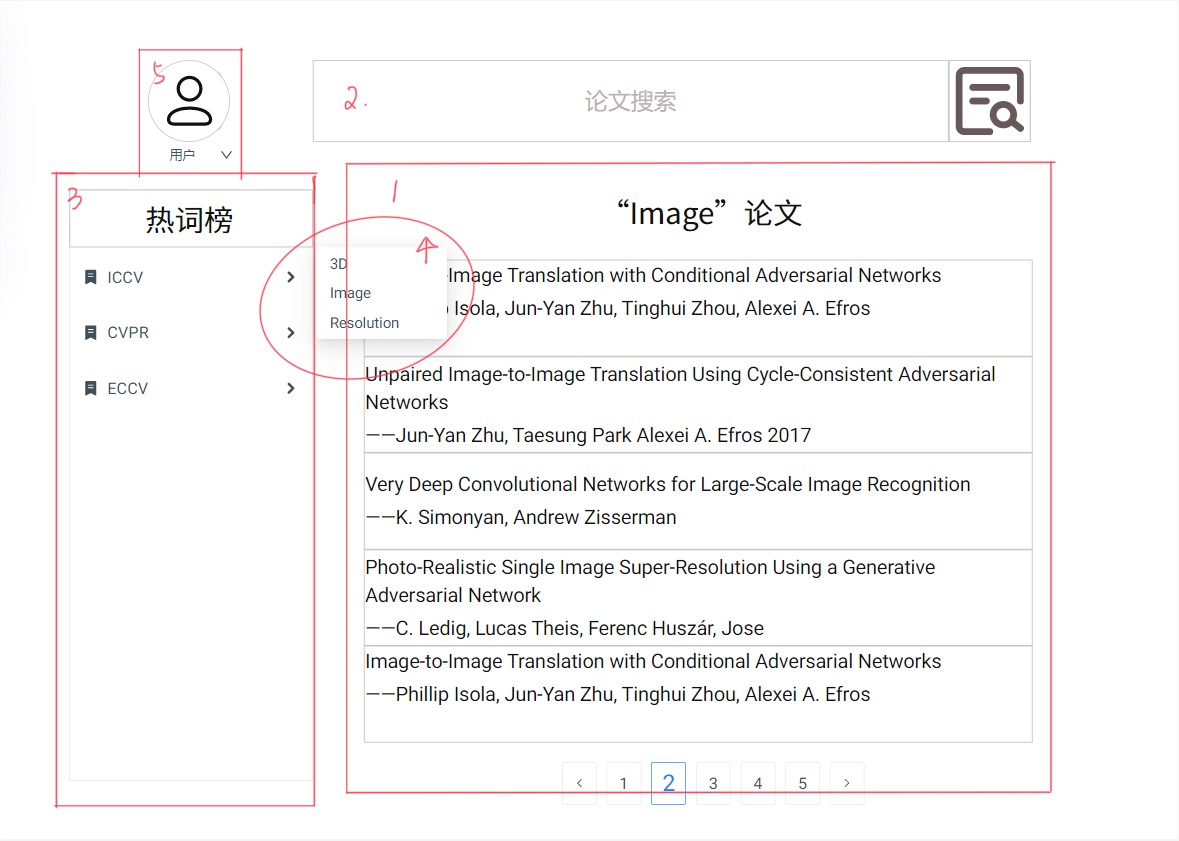


图 用户数据库

## 界面设计





各区域功能如下：

1. 展示与对应热词相关的近期的论文题目
2. 提供论文查找功能，用户可提供论文的题目、关键词或作者
3. 提供指定顶级会议的热词
4. 展示具体会议的近期热词
5. 提供用户信息界面，包括用户的信息，历史搜索记录及针对不同用户的推荐论文列表
6. 展示用户的信息
7. 展示用户的历史搜索记录
8. 为用户提供的推荐论文的题目列表