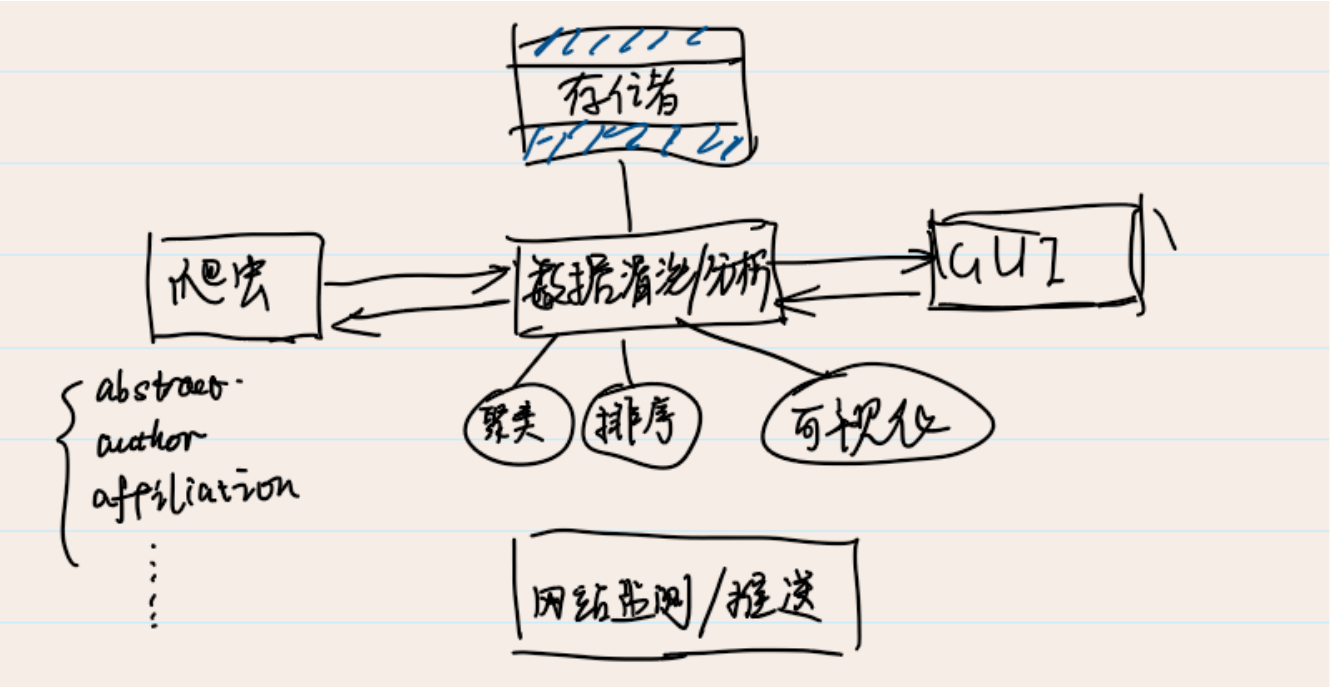
Paper Assistant

总体模块



需求分析

甲方视角

1. 基础需求
    - 关键字查找：例如给定title, author等，返回abstract, comment等
2. 高级需求
    - 统计分析：例如对于给定会议/期刊，分析其热度；

      给定研究机构，绘制其研究领域关键词图等
    - 网站监测与自动推送：当关注网站的相关内容有更新，自动推送。

乙方视角

各个模块需要实现的功能：

- GUI：图形界面
- 数据清洗/分析：

    - 相关性排序
    - 聚类/分类
    - 可视化
    - 喜好分析，评论分析（可选）
- 爬虫：
    - 获取数据。爬取ACM, IEEE, Arxiv等网站获取文章， 爬取twitter, reddit, 微信公众号推送等获取评论
    - 网站监测。监测关注网站的变化。

接口定义

- 爬虫<-->数据清洗/分析

  dictionary

字段	类型	含义
keyword	list	搜索关键字
title	string	论文名
authors	list	作者列表
affiliation	string	发表论文的机构
source	string	文章网站来源
abstract	string	论文的摘要部分
comment	list	媒体用户评论或介绍

- 数据清洗/分析<-->GUI

  GUI -->数据清洗/分析：

  发出命令

字段	类型	含义
type	string	命令类型,例如 搜索，分析
keyword		以下均为用户输入内容，类型及含义同上
title		
authors		
affiliation		
source		

  数据清洗/分析-->GUI

根据GUI的命令返回其所需要的结果。需要实现的命令包括：

1. 基础查询：根据相关性排序后返回dictionary
2. 统计查询：例如收到GUI的要求，对于给定会议/期刊，统计各个子领域的出现频次，并将结果可视化返回。（实现1~2个用于展示，其余的在PPT中写明即可）

任务分配

爬虫/网站监测：李文睿

待实现功能：

- 爬虫。爬取ACM, IEEE, Arxiv等网站获取文章， 爬取twitter, reddit, 微信公众号推送等获取评论
- 网站监测（用于自动推荐）。给定关注的网站及关注的内容，监测关注内容的更新并返回结果。

实现建议：

1. 参考github上已有代码。
2. 第一周先实现爬取一个网站后，与数据分析部分协调好接口。剩余网站第二周实现。

数据清洗/分析：王瑞凯

待实现功能：

- 基础查询：相关性排序
- 统计查询：（举例）对于给定会议/期刊，统计各个子领域的出现频次，并将结果可视化返回。
  （可以自己设计，实现1~2个用于展示，其余的在PPT中写明即可）
- 喜好分析：根据历史查询进行分析，用于推荐。

实现建议：

1. 第一周写好功能设计（该部分的PPT很重要），以及打算写代码的部分
2. 等待爬虫和GUI的接口完全定义好再开始实现（第二周开始）

GUI： 戴路

待实现功能：GUI

实现建议：第一周完成，给数据分析提供接口。