

设计文档

小队成员：

031602121 乐忠豪

031602102 蔡子阳

1. 背景.....	3
2. 设计目标.....	3
1.1 需求分析.....	3
1.2 性能指标.....	3
3. 模块设计.....	4
3.1 模块流程图及说明.....	4
3.2 数据结构说明.....	4
3.3 算法描述.....	4
3.4 与其它模块的接口.....	5
3.5 异常处理.....	5
3.6 测试考虑.....	5
4. 系统集成包装.....	5
5. 设计评审意见.....	5

1. 背景

小樱是一名大三的学生，一直痴迷于吃鸡类游戏，某日听闻同宿舍的小狼刚和导师去参加了 CVPR 会议，内心羡慕不已，便下定决心痛改前非、努力钻研，希望能在毕业前完成一篇站在时代前沿的优秀论文。但令人苦恼的是，他不知道近几年顶会的热门领域和研究方向，根据论文 list 去一篇一篇查找总结效率又着实太低，于是求助于“软工实践互助爱心组织”，希望我们能帮助他设计一个平台解决现阶段的需求。

2. 设计目标

1.1 需求分析

一、用户给定一个论文列表

1. 要求实现通过论文列表爬取论文的题目、摘要以及原文链接。
2. 可实现对论文列表的增删改操作（今年、近两年、近三年）。

二、对爬取的信息进行结构化处理，分析 top10 个热门领域或热门研究方向

1. 可对论文属性（oral、spotlight、poster）进行筛选及分析。
2. 形成如热词图谱之类直观的查看方式。

三、可进行论文检索，当用户输入论文编号、题目、作者等基本信息，分析返回相关的 paper、source code、homepage 等信息。

四、可对多年间、不同顶会的热词呈现热度走势对比（这里将范畴限定在计算机视觉的三大顶会 CVPR、ICCV、ECCV 内）。

五、可进行数据统计，例如每个国家录用文章的分析、每个学校录用文章的分析、哪个学校哪方面的研究方向比较强等。

用户附加需求：

在不改变设计理念、符合用户使用习惯的前提下，在上述需求的基础上进行扩充升级，或发挥想象能力为原型添加自己的 idea。

1.2 性能指标

从需求来看，最主要的时间耗费在爬去论文信息以及存储论文结构体上，用户的目标体量在 9000 篇论文左右（3 年的论文*3 大顶会*每大顶会 1000 篇）。论文题目不超过 200 个

字符，总体量为 200W 字符。

1. 完成全部论文标题遍历及存储需 5s 的响应时间。
2. （爬取时间未知）
3. 检索论文题目及相应信息、形成热词图谱的时间各在 1s 内。
4. 添加功能可以通过论文编号、论文关键词查找相关论文。

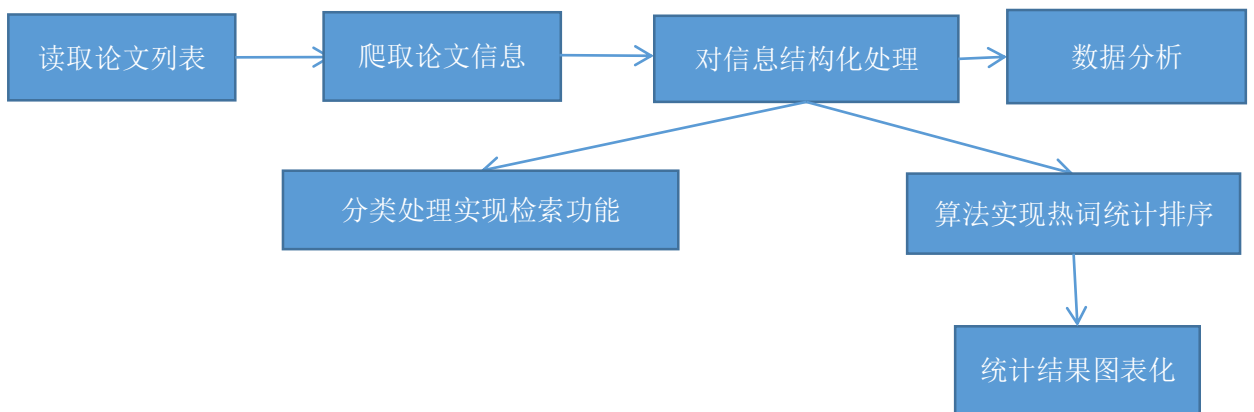
3. 模块设计

3.1 模块流程图及说明

设计实现流程如下：

通过用户给定的论文列表从网页上爬去论文信息（论文编号、作者、原文链接等）

-> 使用 C++数据结构以及 `map<string,class>`容器存储爬去到的论文信息-> 通过 `map` 容器实现数据检索，热词统计（添加通过文章关键词检索相关文章的功能）以及数据分析-> 将统计结果以图像的形式展现出来



3.2 数据结构说明

论文类：

```
{
```

属性：

论文 ID；论文题目；论文摘要；论文年份；论文作者；论文属性（oral、spotlight、poster）；关键词（用作词频统计）；type（表明属于哪一个会议）

成员函数：

获取类属性的各个值；

```
}
```

3.3 算法描述

首先使用 python 编写爬虫内嵌至 C++中，将得到的数据存储于文件中，对文件遍历存储于 `map` 容器中（基于 `key,value` 值存储功能），通过 `map` 容器的红黑二叉树进行数据查询

访问，作出词频统计分析，以及检索筛选算法。

3.4 与其它模块的接口

- 1.爬虫接口
- 2.类接口

3.5 异常处理

1. 读取论文列表失败。
2. 爬取论文信息失败。
3. 读入爬取数据失败。
4. 论文列表中无用户输入的论文标题。

3.6 测试考虑

单元测试：爬虫测试，类读取信息测试，词频统计排序测试

集成测试：测试能否得到预期需求的效果

测试工具：Visual Studio 2017

4.系统集成包装

将系统包装成一个拥有界面的 windows 窗体软件

5.设计评审意见

No.	问题描述	提出人	处理方式/说明	状态
				Open
				Closed