Jelly 精准的大学生校园搜搜引擎学术报告

上海理工大学，刘鹏、崔艺聪、张卉雪、王启帆

张冰雪，讲师

摘要

本项目是为解决大学生校内信息获取途径局限、分散以及信息获取不及时等问题所产生，我们项目的实现形式：一款指向大学生的基于校内生活学习娱乐等方面信息的精准搜索引擎--Jelly应运而生。基准是一个面向大学生的校内信息的搜索引擎，旨在为大学生提供一个校内公共信息与学生经验信息的便捷高效的获取途径，主要定位是精准，精准的指向大学生，精准的提供服务。而实现这种精准服务的最好方式，就是用户驱动的搜索引擎。

关键词：信息获取，及时性，搜索引擎, 用户驱动

背景

作品产生来源于我自己的经历，之前有一次有个较为紧急的事情要咨询辅导员，由于当时不方便只能在网上问他，但是辅导员隔了很久才回我，导致事情有点耽搁。但实际上后面发现我的问题在学校相关网站都有说明，只是不知道如何获取。正好那段时间自己在学习写爬虫，所以我就想能不能做个爬虫专门爬取学校各方面信息，再进行一定的数据清理和整合，基于此建立一个搜索引擎，当学生有什么事项不清楚或需要查询相关通知和个人信息的时候，可以直接通过搜索引擎获得信息

方案论证

本作品的核心是一个搜索引擎，而搜索是一种集成的，跨越时间、网络空间限制直指对应通知信息的一种消息获取方式，完美契合了学生在学校获取信息渠道散乱、主动性不足的问题。既可以学生主动获取自己想要的信息，也能通过学生的个人信息推送相关重要通知，不至于因为信息获取不便的问题而错过相关事务。在项目开始之前，也通过与同学和老师的讨论，以及国内外相似产品的调研，论证了本项目的合理性、功能性和可靠性。

合理性：搜索引擎是一种聚集的、跨越时间、网络空间限制直指对应通知信息的一种消息获取途径。其特性完美解决了上述的大学生信息获取问题。

功能性：它对校园平台中的信息进行分析整理，一方面学生能高效的找到所需信息。一方面我们也对重要信息进行推送

可靠性：为了保证通知的可靠性，我们作品提供的搜索结果实质是一系列通知的Url,学生选取相应的项后，进入学校网站对应的通知页。

研究方法

对于本论文实现所运用的研究方法主要涉及：

1网站设计：

其中涉及的核心方法如下：

1）搜索框架haystack：

该框架能满足绝大部分的站内检索需求，但是对于一个搜索引擎，有很多需要个性化设计和算法设计的地方收到限制，为此我们查看源码后，重写了其中的很多方法，使其能更好的为我们所用。但后期如果要逐渐优化，我们可能需要修改源码。

2） 前端框架purecss

选取框架是我们考虑bootstrap和pure，最终选择Pure的原因是pure设计比较自由、样式比较新颖，并且开发相对轻巧。

3）索引引擎elasticsearch

该索引引擎是基于java的，性能优良，初期我们使用的纯python的索引引擎，但由于速度较慢，转而使用elasticsearch

2数据分析：从上一步潜在信息中分析出信息之间的联系和规律

3机器学习（深度学习）：机器自动学习用户的数据，自动分类。

4自然语言处理：理解搜索引擎中的文本（包括用户输入、通知等），从而能帮助机器更好的学习

5推荐系统：设计个性化的推荐算法，精准有效的向用户推送信息服务

研究结果

本作品有如下功能：

1） 学生根据个人需求主动搜索校内信息。

2） 网站根据学生个人情况及时推送与学生相关的信息。

为实现上述功能，网站需要实现下列模块：

1）爬虫模块

因为网站需要提供学生的校内信息查询，目前我们还没有与学校协商，得到其数据库权限，所以需要利用网络爬虫定时爬取并更新学校各网站的通知（我们主要向学生提供链接），从而向学生提供服务

2）搜索模块

搜索模块主要对爬取的通知信息及下载链接等，提供搜索服务，在网站搜索时用户输入的基础上，再考虑用户的个人信息、通知来源网站、通知发布时间、通知浏览次数等因素，返回匹配的通知链接和相关摘要，用户可通过搜索结果的各个条目进入对应通知页面

3）通知推荐模块

根据学生的专业、年级方面推送相关通知，目的是为了使学生不错过重要信息，

4）学生登录与注册模块

参考文献

1 Zhiguo Wang, Wael Hamza, Radu Florian. Bilateral Multi-Perspective Matching for Natural Language Sentenes,// IJCAI 2017.

2 周志华， 机器学习，北京，清华大学出版社，2017。

3 李华，统计学习方法，北京，清华大学出版社，2017。