自选项目信息描述

# 项目名称与参与成员名单

|  |  |
| --- | --- |
| 项目名 | 专家发现系统 |
|  | |
| 项目组成员 | 1.蒋鑫浩  2.刘振宇  3.涂昌焕  4.周春鹏 |

# 二． 项目相关摘要描述

在知识经济时代的大背景下，政产学研新经济模式被迅速推广，目的在于借助政府的力量，推动企业、高校、科研院所间的合作，为联结科学创新和技术商业化搭建桥梁，目前这一合作模式在世界各国都倍受重视。但国产业核心技术仍严重依赖国外，高校和科研机构的科技成果转化率偏低，人才培养与研发实践脱节，政府、高校、企业和科研机构均无法发挥各自的能力优势，因此形成了企业渴求专家人才，而专家人才无法为企业效力的困境。目前，各地政府解决这一问题的主要途径依赖于人工方式，如通过相关研究领域的负责人介绍和推荐，少部分地区为促进产学研合作的计划，采取企业资助大学搞科研、鼓励大学参加企业科研、建设科技园和创新中心、组织人才培养和交流合作、建立教育捐赠的财会制度等措施，但能寻找到专家数量少，专业领域有限，周期长，资源耗费开销大，工作效率低下，使得政用产学研的推进进程艰难缓慢。因此搜索专家信息，发现专家，合理利用配置专家资源是当前亟待解决的重要问题。总体而言，我们的项目系统旨在解决企业-专家学者对接困难的问题，从而提高产学研对接效率。

# 三． 项目需求描述

满足政府和高校学者的对接，实现产学研相结合，满足政府对某一领域的专家学者的需求。

3.1 功能性业务需求

爬取高校相关科研团队的信息；

根据高校的专家学者的信息生成关于研究领域及其相关技术的用户画像；

展示专家的已有信息及研究成果，如科研论文、专利等；

利用数据挖掘技术，根据科研成果建立起科研人员之间的关联；

根据提供的政府工作人员数据，生成相应的用户画像；

在政府工作人员的用户画像和专家学者的用户画像之间建立联系，为相关度较高的双方进行推荐。

3.2 非功能性业务需求

1. 性能需求：相应速度较快，结果精确，消耗资源少；

2. 可靠性需求：系统较为稳定，故障可预测性高；

3. 易用性需求：系统界面简洁，操作简单，交互性高；

4. 运行环境需求：系统运行环境要求低，消耗硬件资源较少；

# 四． 项目功能结构

本项目主要分为3个模块：信息展现模块、信息检索模块、专家信息可视化模块、

1. 信息展现模块：
2. 专家画像：可以展现知名专家的基本信息、任职生涯、研究方向，合作团队，发明专利，科研项目等信息。
3. 论文信息：可以展现国内各类研究方向论文的基本信息，发表作者，发表时间，研究方向。
4. 研究团队：可以展现国内各类专家形成的研究团队的基本信息，人员组成、专家属性、研究主题、科研统计等。
5. 信息检索模块：
6. 专家搜索模块：可以根据用户搜索词，匹配主题关键词并对主题内专家进行搜索，最后将搜索结果按照匹配度进行排序。
7. 专家信息可视化模块：
8. 按照全国层面-区域/省层面-高校/学院层面进行划分，对分析得到的各项信息进行数据可视化。
9. 按照专家的研究团队组成，合著关系，构建特定专家的自我网络，并且将专家的自我网络进行数据的可视化。

# 五. 项目预期使用的技术

Python数据爬取技术。

Scrapy爬虫框架，适合爬去大量资源。

Flask，We应用程序框架。

Bokeh，数据可视化技术。

MySQL，数据库。

Neo4j，NoSQL图形数据库。

MongoDB，分布式文件存储数据库。

Faiss，一个用于相似性搜索和密集向量聚类的高性能库。

BERT，一种以Transformers为主要框架的双向编码表征模型。

技术难点与风险。

本项目要用到的很多技术项目小组的成员都没有接触过，甚至有的比较小众的软件框架以前都没有听说过。因此，项目团队可能会因为对技术的不熟悉而影响项目的成功。例如：培训的不足而导致对技术的理解不够深入、技术应用上的经验不足而导致项目进度的延期等。

项目中的专家数据信息来源于各大高校的主页，但是不同高校的网页的结构千差万别，无法用一个通用的爬虫来爬取需要的专家数据，而如果采用针对特定网页的特定的爬虫策略来爬取的话，所耗费的工作量将是极大的，在短时间内完成的可能性很小。