项目报告

第一部分：研发报告

1. 功能：开发的项目每个模块包括什么，实现了什么
2. 程序架构图



图1.程序框图

程序框架图如图1所示，每个步骤描述如下：

1. 官方网站（中国科学院、中国工程院）抓取、挖掘，得到院士链表包括院士姓名、单位、研究方向、出生年月、获评院士年份等；
2. 在“发表文章数据库”检索一名院士的所有文章，把院士本人及所有合作作者作为一个新链表，合作研究者链表，包括研究者的姓名、单位；
3. 把合作研究者链表的每一个研究者在“发表文章数据库”检索，得到正确的姓名-单位关系，并得到文章-研究者关系库；
4. 分析文章-研究者关系库中的关联关系，计算出研究者链表中每两名研究者之间的合作关系，得到合作关系矩阵，至此可以得到院士关系网络中，任意两人之间的“亲密度”；
5. 考虑到单位的重要性，将研究者-研究者之间的单位关系数值化后形成单位关系矩阵；
6. 合作关系矩阵和单位关系矩阵相乘，得到关系矩阵；
7. 通过关系矩阵计算研究者的重要性，并按照重要性进行排序；
8. 以此院士为树干，开始按照前面的排序安放每个研究者在关系树中的位置，安放原则为：1）关系矩阵数值较大，2）若矩阵数值相等，安放在远离树干的地方，与此同时，计算该研究者父辈节点，和院士的距离，以及有多少子节点；
9. 下一个院士，循环b-h。
10. 关键技术：遇到难点，关键算法，遗留问题
11. “发表文章数据库”，每一篇文章为一条记录，作者姓名和单位存在多对多的情况。解决算法，将研究者姓名-研究单位放入“发表文章数据库”检索，检索到的文章数量作为判定单位正确与否的标准；
12. 文章的单位存在数据格式不规范情况（包括：单位的表达并不规范，同一研究者单位不同表达）。解决算法，1，单位的抓取采用多种方法，HttpWebResponse，HtmlAgilityPack交叉使用得到更为精准的单位表达，排除错误信息，2，若同一姓名，单位存在包含或被包含关系，视为同一作者；
13. 在上述算法中，关系图谱对研究者的排序依赖性很强，即排序决定了关系图谱的地位，有时出现正在执行的院士不是排在第一的位置，则出现研究者和他脱离的现象,解决方法，将院士脱离出来，赋予更高权限；
14. 遗留问题：1、同院士相关的研究者只有一个主单位，目前算法不能实现研究者多个单位的处理；2、算法需要检测、挖掘、存储的数据量过大，算法需要优化。
15. 文件：物理存储及程序运行

数据爬取：

1. 使用java 开发环境，利用jsoup包，对中国科学院、中国工程院网站上的院士信息进行抓取、存储。
2. 使用C#开发环境，设计WindowsFormsApplication程序界面，使用WebBrowser类检索、抓取“发表文章数据库”，辅助HttpWebResponse抓取数据，得到合作研究者的姓名、单位、发表文章等信息，并上传至数据库。

数据挖掘：

1. 使用C#开发环境，设计控制台程序，对数据进行挖掘、处理、存储。

第二部分：数据库使用手册

1. 数据库结构图
2. 数据内容说明
3. 数据库表解释
4. Dbpaper. Academincian 院士基本信息

|  |  |  |
| --- | --- | --- |
| Academincian | 定义/内容 | 说明 |
| name | 姓名 |  |
| url | 院士简介url |  |
| detail1 | 简介1 | 来源url，经历 |
| detail2 | 简介2 | 来源url，研究方向 |
| acayear | 获选院士时间 | 正则表达式 |
| birthday | 出生年月 | 正则表达式 |
| departmentauto | 第一单位 | 正则表达式计算 |
| Imageurl | 照片url | 来源url |
| ancestors | 籍贯 |  |

1. Dbpaper. Relationtoaca\_copy 有关研究者与院士关系

|  |  |  |
| --- | --- | --- |
| Dbpaper. Relationtoaca\_copy | 定义/内容 | 说明 |
| id |  |  |
| academincian | 院士姓名 |  |
| name | 研究者姓名 |  |
| department | 研究者单位 | 算法计算来 |
| level1\_number | 研究者与院士共同发表文章数 |  |
| level2\_number | 研究者CNKI上发表文章数量 |  |
| link\_name | 研究者的上级研究者姓名 | 算法计算 |
| link\_department | 上级研究者单位 |  |
| startpublish | 发表文章最早时间 |  |
| endpublish | 发表文章结束时间 |  |
| keyword1 | 主研究方向 |  |
| Keyword2 | 次研究方向 |  |
| isAcademincian | 是否是院士 | 姓名-单位和院士表一致 |
| childrenNumber | Tree上位置有多少子节点 |  |
| relationLevel | Tree上距离院士的距离 |  |
| Flag | 是否新更新的 | 算法修正 |

1. Dbpaper. Paper\_now 文章存放

|  |  |  |
| --- | --- | --- |
| Paper\_now | 定义/内容 | 说明 |
| paper\_id | 文章id |  |
| url | 文章的url |  |
| keywords | 文章关键词 |  |
| title | 文章题目 |  |
| publishyear | 发表年 |  |
| keywords | 关键词 |  |
| authors | 作者姓名 |  |
| academician | 院士姓名 |  |
| title | 题目 |  |
| categorytag | 分类 |  |
| citednum | 引用数量 |  |
| funding | 基金 |  |
| downloadnum | 下载数量 |  |
|  |  |  |

1. Dbpaper. ReasearcherLink合作文章数量

|  |  |  |
| --- | --- | --- |
| ReasearcherLink | 定义/内容 | 说明 |
| idLink | 链接id |  |
| startResearcherId | 作者1id | (Relationtoaca.id) |
| endResearcherId | 作者2id | (Relationtoaca.id) |
| value1 | 合作论文数 |  |
| academincianId | 隶属院士id | ( Academincian.id) |

第三部分：可视化展示

1. 图
2. 图解释

第四部分：测试报告

1. 测试用例
2. 测试结果