一、前期工作总结

1. 主要完成工作

对于网络上的海量网页，设计并实现针对政治人物、高校教师领域的人员简历识别系统。通过预先构建的过滤词表对非人员简历的网页进行初步过滤，提高识别效率。将简历识别问题看做分类问题，根据人员简历内容包含的特点，提出基于启发式规则和机器学习相结合的分类方法。

2. 训练数据集

正例：“中国共产党新闻网”中国领导干部资料库 4095篇

中国科学院大学“硕士生导师”的简历 6860篇

负例：搜狗新闻 9960篇

3. 总体思路

1. 构建过滤词表（“启事”，“申报”，“通知”，“讲话”，“招聘”等共59个词条），对采集到的网页文本首先进行初步过滤。
2. 构建8个词典：学术组织，研究方向，简历子标题词，简历高频词，职务及职称，民族，“简历”的同义词，匹配邮箱、网址、电话号码、专利号等内容的正则表达式。
3. 对训练数据集抽取以下15个特征：

表1-1 SVM分类特征

|  |  |
| --- | --- |
| 序号 | 特征内容 |
| f1 | 文本长度 |
| f2 | 包含邮箱的数量 |
| f3 | 包含电话号码的数量 |
| f4 | 包含网址的数量 |
| f5 | 包含专利号的数量 |
| f6 | 包含专著、论文的数量 |
| f7 | 包含书名号的数量 |
| f8 | 包含年份词的数量 |
| f9 | 包含学术组织词条的数量 |
| f10 | 包含研究方向词条的数量 |
| f11 | 包含简历子标题词的数量 |
| f12 | 包含简历高频词的数量 |
| f13 | 包含职务、职称词的数量 |
| f14 | 是否包含民族词条 |
| f15 | 是否含有‘简历’或‘简历’的同义词 |

并处理成libsvm格式，利用SVM训练分类器。

1. 利用过滤词表对测试数据集进行初步过滤。之后抽取特征，处理成libsvm格式后，利用上一步生成的SVM分类器进行分类判断即可。

4. 测试结果

1. 北京大学环境科学与工程学院网站：1020个页面

准确率：97.29%（36/37） 召回率：70.58%（36/51）

1. 上海交通大学环境科学与工程学院网站：599个页面

准确率：100%（43/43） 召回率：78.18%（43/55）

1. 济南市人民政府网站：1270个页面

准确率：100%（8/8） 召回率：88.88%（8/9）

1. 吉安市人民政府网站：2356个页面

准确率：95.23%（20/21） 召回率：74.07%（20/27）