

1 绪论

1.1 邮件分类技术的现状与背景

随着互联网和 IT 产业的高速发展,我国网民的规模一直不断增长,人民的生活和工作已经离不开网络了。中国互联网络信息中心 (CNNIC) 于 2019 年 8 月 30 日在北京发布了第 44 次《中国互联网络发展状况统计报告》^[1], 报告显示截至 2019 年 6 月,我国网民规模较 2018 年底增长 2598 万,达 8.54 亿。互联网已经覆盖了全年龄段用户,各年龄段用户增长都呈现出较快增长。随着互联网的蓬勃发展,电子邮件因其能够迅捷地与他人交流在生活和工作中扮演相当重要的角色,特别是在保障个人账号安全的方面。与此同时,随着大数据时代的到来,个人邮箱账号很容易被泄露,我们的邮箱经常会收到大量垃圾邮件,其中不乏广告诈骗和钓鱼邮件。来自于中国互联网互联 (ISC) 发布的《2013 年第三季度中国反垃圾邮件状况调查报告》^[2]的权威数据表明,在中国平均每个邮箱账号每周会收到垃圾邮件 11.9 封,占每周收到总到总邮件 32.8%。尽管此次统计垃圾邮件在所有邮件中整体占比较往年有所所有下降,但“金融保险类”和“互联网服务类”涉及个人金融安全的邮件占比正快速增长。在此背景之下,如何在较短时间内分辨出正常邮件和垃圾邮件对于提高个人工作效率、保护个人隐私和财产安全十分重要。

国内外许多学者对邮件分类做了许多深入研究,其中包括朴素贝叶斯分类算法和 K 近邻算法,以及文本分类方法和结合用户属性的方法,以下是几种常见的算法^[3]。

(1) SVM 算法。支持向量机 (support vector machine, SVM) 是一种基于统计学习理论的模式识别方法,目前被经常大量广泛地应用于生物信息学、文本分类等多个方面^[5]。SVM 算法被提出后快速发展,在许多应用领域都得到了广泛应用。它的原理如下:第一步通过使用核函数把低维线性不可分的数据转化到高维,实现线性分割。第二步,利用最大化间隔得到间隔最大的分割线,得出支持向量。第三步利用得出的模型进行分类的预测。

(2) 朴素贝叶斯分类算法。通过在贝叶斯算法的基础上进行了相应的简化处理,最终得到朴素贝叶斯算法。假定预先设定目标值时属性之间都是相互条件独立的,任何一个属性变量都不会对最终决策结果占有着大的比重,也不会有哪个属性变量对最终决策结果有着较小的比重。这个简化思想虽然降低了贝叶斯分类算法在实际应用中的分类效果,但简化了贝叶斯方法需要考虑的因素。在理论上和其他算法比较而言

，朴素贝叶斯模型的正确率往往是最高的。但实际上这种结果也是有误差的，因为这是在属性彼此独立的前提下进行的。在实际应用中朴素贝叶斯算法中各属性通常是不独立的。属性间没有独立性或独立性较低以及属性数量较多时，分类的效果往往会出现很大误差。如果属性个数较少或相关性很小，朴素贝叶斯算法的性能将是最佳的。

(3) K 近邻算法。K 最近邻(k - Nearest Neighbor)分类算法，这种方法的思想是假设在特征空间中，一个样本附近的 k 个最近的(即特征空间中最相邻的)样本大多数属于某一个类型则可以判断测试样本也属于这个类型^[4]。K 近邻算法是机器学习中常见的快捷方法，也是一个在实际中经常使用的算法。KNN 算法缺少训练过程，虽然经常取得令人期望的结果，但是由于其复杂的计算过程会耗费较长的时间，所以该算法对计算速度要求较高的垃圾邮件分类领域并不适用。

1.2 本论文的结构

完成本次毕业设计需要使用到 Anaconda Python 集成开发软件。第一步需要使用 Anaconda 平台下的 Jupyter Notebook 软件进行朴素贝叶斯算法仿真。在算法仿真成功后，说明使用朴素贝叶斯算法处理邮件分类是可行的。接下来准备一些可供程序学习的样本和测试文件，打开 Spyder，完成 Python 代码的编写后导入样本以供程序自我学习，导入新的测试邮件最后输出识别分类的结果。

本论文主要由五个章节构成。第一章节是绪论部分，主要简单说明了研究的背景和当前邮件分类技术的发展现状，以及项目的过程和结构；第二章节介绍了项目过程使用的开发技术、工具和环境，详细阐明了本次设计将要用到的软件平台、语言以及各自的特点；第三个章节是算法的研究与仿真，首先要论证算法的可行性，之后用 Jupyter Notebook 软件仿真算法，验证算法的正确性；第四部分是软件的实现，详细阐述了软件的运行和在 Spyder 中完成代码的编写；第五部分是运行结果，用于展示最终程序运行的正确性。

2 开发环境以及开发工具介绍

完成本次毕业设计运用了 Anaconda 集成软件完成 Python 代码的编写，以下是对相关软件和语言及其各自特点的简要介绍。

2.1 Anaconda

这是一个是将 Python 以及基于 Python 的用于科学计算所需的扩展库编译打包好，并且简化模块和环境的管理，同时包含多个编译程序的发行版本。Anaconda 包含了多个 Python 代码编译软件以及高达近 200 个科学包及其依赖项，Anaconda 的特点如下^[5]：

- (1) 包含了众多当下热门和前沿的 Python 包，例如科学、数据分析、数学、工程等。
- (2) 完全免费、开源，只有额外的加速、优化是需要付费的。
- (3) 免费的社区支持。
- (4) 全平台支持：可应用于 Windows、Mac、Linux 等各个操作系统；支持多个版本的 Python，可自由切换。

Anaconda 的主要应用如下。

Anaconda Navigator。最主要的应用程序，用于工具包和编译软件的管理，能够实现图形可视化的下载、更新和删除库和环境。

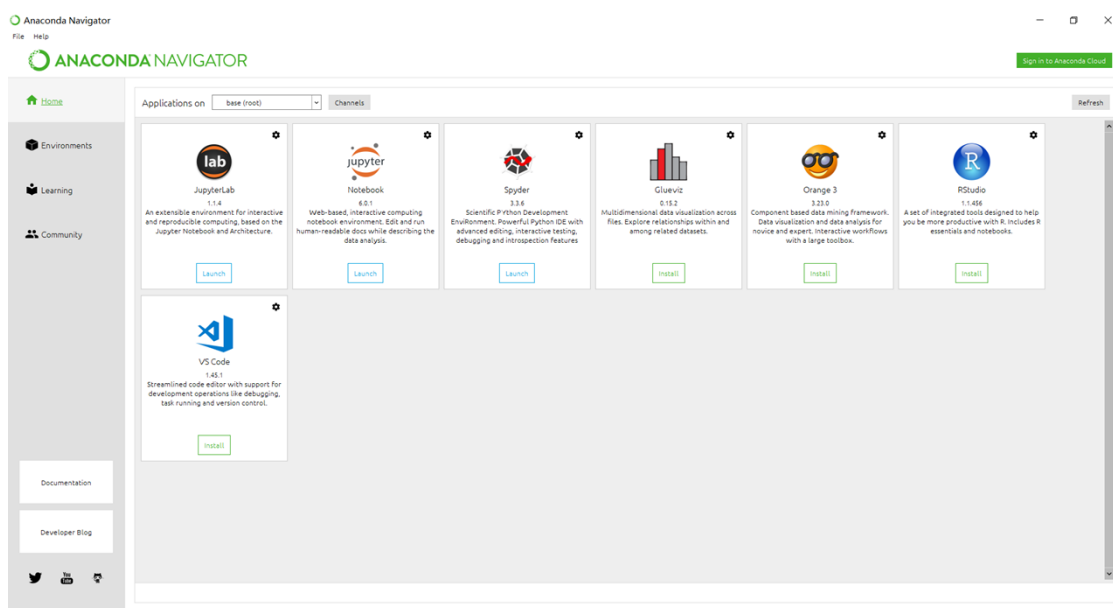


图 2.1 Anaconda Navigator 界面

以上内容仅为本文档的试下载部分，为可阅读页数的一半内容。如要
下载或阅读全文，请访问：

<https://d.book118.com/957161066016006141>