2022 年全国大学生信息安全竞赛

作品报告

作品名称：PoliScope：动静混合分析的 APP 隐私合规检测

电子邮箱：shenaowang@foxmail.com

提交日期：2022.05.07

填写说明

1. 所有参赛项目必须为一个基本完整的设计。作品报告书旨在能够清晰准确地阐述（或图示）该参赛队的参赛项目（或方案）。

2. 作品报告采用 A4 纸撰写。除标题外，所有内容必需为宋体、小四号字、1.5 倍行距。

3. 作品报告中各项目说明文字部分仅供参考，作品报告书撰写完毕后，请删除所有说明文字。(本页不删除)

4. 作品报告模板里已经列的内容仅供参考，作者可以在此基础上增加内容或对文档结构进行微调。

5. 为保证网评的公平、公正，作品报告中应避免出现作者所在学校、院系和指导教师等泄露身份的信息。

目录

摘要

在新冠肺炎疫情影响下，远程办公短期需求爆发，巨大的需求瞬间引爆了办公平台市场。据百度热度指数，2020 年 2 月，远程办公搜索指数同比增长 491%，环比 1 月份增长 317%。远程办公平台的主要功能是通过第三方插件、软件、网站等工具，实现传统办公地点范围之外进行协同工作。远程办公平台中其中往往集成了大量的第三方应用。但是利用 Web API 对服务器进行请求时，出现的各种集成服务在功能和流程上通常会变得更加复杂，这种复杂性使得原有的应用程序与远程办公平台在边界交互产生了新的安全威胁，所以关于远程办公平台的 API 的安全威胁不容小觑。

虽然近些年来国内外已经有较多关于 API 的研究，但专门针对远程办公平台 API 安全问题的研究却是少之又少，很多时候只有当漏洞出现甚至引发较大事故时才会发现平台存在的安全问题，因此有必要设计一个自动化检测工具，对现有远程办公平台的 API 进行安全性排查，及时发现并修复问题，把相关安全风险降到最低。

本作品基于 Beautiful Soup 库和 Selenium 框架实现了对远程办公平台 API 文档信息的自动化收集，使用自然语言处理技术（NLP），通过使用 spaCy 框架（根据依存分析）将 API 进行处理与分类并确定其依赖关系，基于 API 调用过程中的依赖关系自动生成请求序列，依照 API 调用的安全规范，准确、快速地对远程办公平台大量 API 服务的安全性进行检测。只要平台提供了为第三方开发提供了 API 接口，API Guardian 就可以对其进行规范分析和安全检测。我们利用该作品在 Slack，Microsoft Teams，Facebook Workspace，Cisco Webex，Discord 等一系列平台检测出相关漏洞 11 个，并向相应厂商提供了漏洞检测报告。

本作品着眼于远程办公平台的 API 安全问题，采用基于细粒度的方法对远程办公平台第三方应用开发过程中的 API 安全规范进行轻量级分析，涉及平台安全，通信安全和应用安全三方面，不仅研究领域新颖，并且采取开拓性的思路，具有广泛的实用价值。

关键字：API 安全　远程办公平台　NLP　漏洞检测

Abstract

Under the influence of the COVID-19 pandemic, short-term demand for remote work broke out, and the huge demand instantly detonated the office platform market. According to Baidu's popularity index, in February 2020, the remote office search index increased by 491% year-on-year and 317% month-on-month. The main function of the remote office platform is to realize non-local working through third-party plug-ins, software, websites and other tools. Among them, a large number of third-party applications are integrated, but when using Web API to make requests to the server, the various integrated services that appear usually become more complicated in function and flow. This complexity has created new security threats, so the security threats to the API of the remote office platform should not be underestimated.

Although there have been many researches on APIs at home and abroad in recent years, but only few researches emphasized on the API security of remote work platforms. In many cases, the existence of the platforms' vulnerability is only discovered when a major accident occurs. Therefore, it is necessary to design an automated detection tool to check the security of the API of the existing remote work platforms, to find and repair existing problems in time, and minimize related security risks.

Our work is based on the Beautiful Soup library and the Selenium framework to automatically collect the API document of remote work platforms. By using natural language processing technology (NLP) and the spaCy framework (based on dependency analysis), we can process and classify the API and identify there dependency relationship. Based on the dependency relationship in the API call process, the request sequence is automatically generated. We have detect 11 related vulnerabilities by our work, on a series of platforms such as Slack, Microsoft Teams, Facebook Workspace, Cisco Webex, Discord, etc., and provided vulnerability detection reports to the corresponding vendors.

Our work focuses on the API security issues of the remote work platform, and uses a fine-grained method to conduct a lightweight analysis of the API security specifications in the third-party application development process of the remote office platform. It involves platform security, communication security and application security.

Keywords: API security remote work platforms NLP vulnerability detection

第一章 作品概述

1.1 研究背景

随着互联网和信息技术的快速发展和普及，信息通讯的便捷迅速，远程办公这种新型工作方式开始出现。传统的一些工作模式也逐渐被由互联网主导的远程办公的新型模式所取代。人们的工作场所不再局限于办公室，在家办公、异地办公、移动办公等非本地办公形式也得以实现。因远程办公的诸多优点——电子化、分离性和灵活性，远程办公软件受到了很多企业的青睐。据中国软件网 2020 年发布的相关数据显示，2014—2020 年中国远程办公的市场规模逐年上升，在 2020 年中国远程办公市场规模将以 104.27% 的增长率增长到 478 亿人民币。

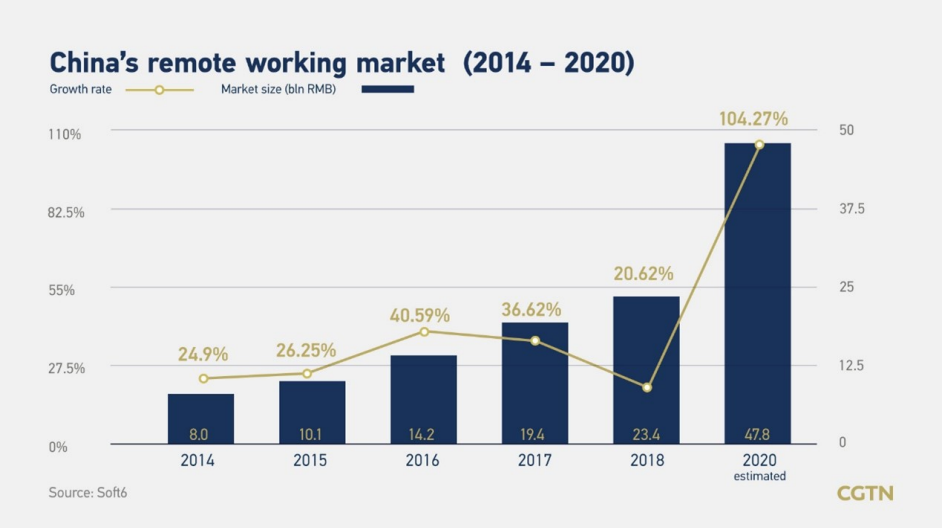


图 1　中国软件网于 2020 年发布的中国 2014—2020 年远程办公市场规模

在新冠肺炎疫情影响下，远程办公短期需求爆发，巨大的需求瞬间引爆了办公平台市场。据百度热度指数，2020 年 2 月，远程办公搜索指数同比增长 491%，环比 1 月份增长 317% [1]。根据 OKTA 于 2020 年 2 月发布的远程办公软件使用用户增长情况显示 [2]，远程办公软件的使用人数因新冠肺炎疫情的影响都大幅提升，其中 Zoom 使用人数同比增长 110

第二章 模板的基本使用

要使用 LaTeX 来完成建模论文，首先要确保正确安装一个 LaTeX 的发行版本。

- Mac 下可以使用 MacTEX
- Linux 下可以使用 TEXLive；
- windows 下可以使用 TEXLive 或者 MikTEX；

具体安装可以参考 Install-LaTeX-Guide-zh-cn 或者其它靠谱的文章。另外可以安装一个易用的编辑器，例如 TEXstudio 。

使用该模板前，请阅读模板的使用说明文档。下面给出模板使用的大概样式。

```
\documentclass{cumcmthesis}
%\documentclass[withoutpreface,bwprint]{cumcmthesis} %去掉封面与编号页

\title{论文题目}
\tihao{A}         % 题号
\baominghao{4321} % 报名号
\schoolname{你的大学}
\membera{成员A}
\memberb{成员B}
\memberc{成员C}
\supervisor{指导老师}
\yearinput{2017}  % 年
\monthinput{08}   % 月
\dayinput{22}     % 日

\begin{document}
    \maketitle
    \begin{abstract}
        摘要的具体内容。
        \keywords{关键词1\quad 关键词2\quad 关键词3}
    \end{abstract}
    \tableofcontents
    \section{问题重述}
    \subsection{问题的提出}
    \section{模型的假设}
    \section{符号说明}
    \begin{center}
        \begin{tabular}{cc}
            \hline
            \makebox[0.3\textwidth][c]{符号} & \makebox[0.4\textwidth][c]{意义} \\ \hline
            D      & 木条宽度（cm） \\ \hline
        \end{tabular}
    \end{center}
    \section{问题分析}
```

```
    \section{总结}
    \begin{thebibliography}{9}%宽度9
        \bibitem{bib:one} ....
    \end{thebibliography}
    \begin{appendices}
        附录的内容。
    \end{appendices}
\end{document}
```

根据要求，电子版论文提交时需去掉封面和编号页。可以加上 withoutpreface 选项来实现，即：

```
\documentclass[withoutpreface]{cumcmthesis}
```

这样就能实现了。打印的时候有超链接的地方不需要彩色，可以加上 bwprint 选项。

另外目录也是不需要的，将 \tableofcontents 注释或删除，目录就不会出现了。

团队的信息填入指定的位置，并且确保信息的正确性，以免因此白忙一场。

编译记得使用 xelatex，而不是用 pdflatex。在命令行编译的可以按如下方式编译：

```
xelatex example
```

或者使用 latexmk 来编译，更推荐这种方式。

```
latexmk -xelatex example
```

下面给出写作与排版上的一些建议。

第三章 图片

建模中不可避免要插入图片。图片可以分为矢量图与位图。位图推荐使用 jpg,png 这两种格式，避免使用 bmp 这类图片，容易出现图片插入失败这样情况的发生。矢量图一般有 pdf,eps ，推荐使用 pdf 格式的图片，尽量不要使用 eps 图片，理由相同。

注意图片的命名，避免使用中文来命名图片，可以用英文与数字的组合来命名图片。避免使用1,2,3 这样顺序的图片命名方式。图片多了，自己都不清楚那张图是什么了，命名尽量让它有意义。下面是一个插图的示例代码。

注意 figure 环境是一个浮动体环境，图片的最终位置可能会跑动。[!h] 中的 h 是 here 的意思，！表示忽略一些浮动体的严格规则。另外里面还可以加上 btp 选项，

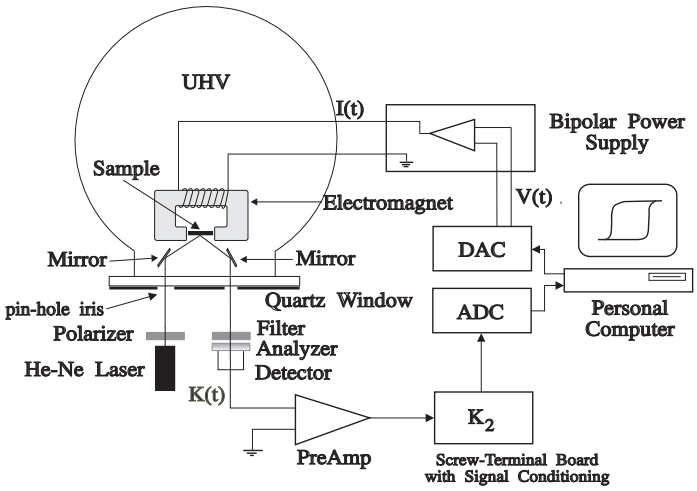
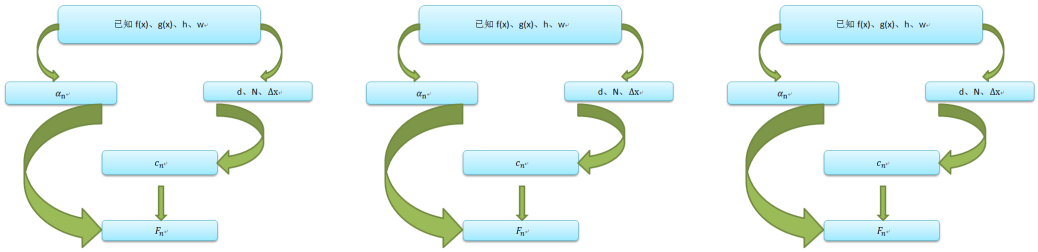


图 2　电路图



(a) 流程图　　**(b) 流程图**　　**(c) 流程图**

图 3　多图并排示例

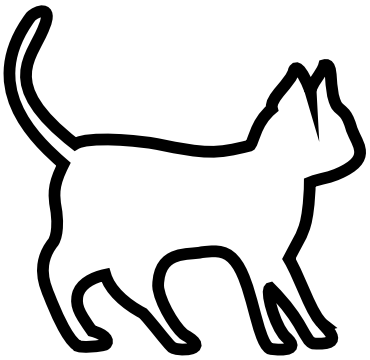
它们分别是 bottom, top, page 的意思。只要这几个参数在花括号里面，作用是不分先后顺序的。page 在这里表示浮动页。

`\label{fig:circuit-diagram}` 是一个标签，供交叉引用使用的。例如引用图片 `\cref{fig:circuit-diagram}` 的实际效果是图 2。图片是自动编号的，比起手动编号，它更加高效。`\cref{label}` 由 `cleveref` 宏包提供，比普通的 `\ref{label}` 更加自动化。`label` 要确保唯一，命名方式推荐用图片的命名方式。

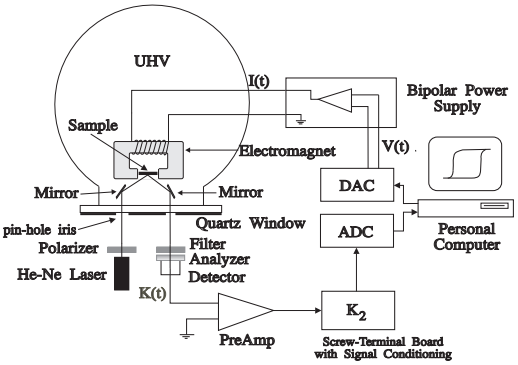
图片并排的需求解决方式多种多样，下面用 `minipage` 环境来展示一个简单的例子。注意，以下例子用到了 `subcaption` 命令，需要加载 subcaption 宏包。

这相当于整体是一张大图片，大图片引用是图 3，子图引用别分是图 3a、图 3b、图 3c。

如果原本两张图片的高度不同，但是希望它们缩放后等高的排在同一行，参考这个例子：



(a) 一只猫



(b) 电路图

图 4　多图并排示例

第四章 绘制普通三线表格

表格应具有三线表格式，因此常用 booktabs 宏包，其标准格式如表 1 所示。

表 1　标准三线表格

D(in)	P_u(lbs)	u_u(in)	β	G_f(psi.in)
5	269.8	0.000674	1.79	0.04089
10	421.0	0.001035	3.59	0.04089
20	640.2	0.001565	7.18	0.04089

其绘制表格的代码及其说明如下。

```
\begin{table}[!htbp]
   \caption[标签名]{中文标题}
   \begin{tabular}{cc...c}
      \toprule[1.5pt]
      表头第1个格 & 表头第2个格 & ... & 表头第n个格 \\
      \midrule[1pt]
      表中数据(1,1) & 表中数据(1,2) & ... & 表中数据(1,n)\\
      表中数据(2,1) & 表中数据(2,2) & ... & 表中数据(2,n)\\
      ..................................................\\
      表中数据(m,1) & 表中数据(m,2) & ... & 表中数据(m,n)\\
      \bottomrule[1.5pt]
   \end{tabular}
```

```
\end{table}
```

table 环境是一个将表格嵌入文本的浮动环境。tabular 环境的必选参数由每列对应一个格式字符所组成：c 表示居中，l 表示左对齐，r 表示右对齐，其总个数应与表的列数相同。此外，`@{文本}`可以出现在任意两个上述的列格式之间，其中的文本将被插入每一行的同一位置。表格的各行以\\分隔，同一行的各列则以 & 分隔。`\toprule`、`\midrule` 和 `\bottomrule` 三个命令是由 booktabs 宏包提供的，其中 `\toprule` 和 `\bottomrule` 分别用来绘制表格的第一条（表格最顶部）和第三条（表格最底部）水平线，`\midrule` 用来绘制第二条（表头之下）水平线，且第一条和第三条水平线的线宽为 1.5pt ，第二条水平线的线宽为 1pt 。引用方法与图片的相同。

第五章 公式

数学建模必然涉及不少数学公式的使用。下面简单介绍一个可能用得上的数学环境。

首先是行内公式，例如 θ 是角度。行内公式使用 `$ $` 包裹。

行间公式不需要编号的可以使用 `\[ \]` 包裹，例如

$$E = mc^2$$

其中 E 是能量，m 是质量，c 是光速。

如果希望某个公式带编号，并且在后文中引用可以参考下面的写法：

$$E = mc^2 \tag{1}$$

式 (1) 是质能方程。

多行公式有时候希望能够在特定的位置对齐，以下是其中一种处理方法。

$$\begin{aligned} P &= UI && (2)\\ &= I^2R && (3) \end{aligned}$$

`&` 是对齐的位置，`&` 可以有多个，但是每行的个数要相同。

矩阵的输入也不难。

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nn} \end{pmatrix}$$

分段函数这些可以用 case 环境，但是它要放在数学环境里面。

$$f(x)=\begin{cases}0 & x\text{为无理数,}\\ 1 & x\text{为有理数.}\end{cases}$$

在数学环境里面，字体用的是数学字体，一般与正文字体不同。假如要公式里面有个别文字，则需要把这部分放在 text 环境里面，即 \text{文本环境} 。

公式中个别需要加粗的字母可以用 $\bm{math symbol}$ 。如 $\alpha a\bm{\alpha a}$ 。

以上仅简单介绍了基础的使用，对于更复杂的需求，可以阅读相关的宏包手册，如 amsmath。

希腊字母这些如果不熟悉，可以去查找符号文件 symbols-a4.pdf ，也可以去 detexify 网站手写识别。另外还有数学公式识别软件 mathpix 。

下面简单介绍一下定理、证明等环境的使用。

定义 1 定义环境

定义 1除了告诉你怎么使用这个环境以外，没有什么其它的意义。

除了 definition 环境，还可以使用 theorem 、lemma、corollary、assumption、conjecture、axiom、principle、problem、example、proof、solution 这些环境，根据论文的实际需求合理使用。

定理 1 这是一个定理。

由定理 1我们知道了定理环境的使用。

引理 1 这是一个引理。

由引理 1我们知道了引理环境的使用。

推论 1 这是一个推论。

由推论 1我们知道了推论环境的使用。

假设 1 这是一个假设。

由假设 1我们知道了假设环境的使用。

猜想 1 这是一个猜想。

由猜想 1我们知道了猜想环境的使用。

公理 1 这是一个公理。

由公理 1我们知道了公理环境的使用。

定律 1 这是一个定律。

由定律 1我们知道了定律环境的使用。

问题 1 这是一个问题。

由问题 1我们知道了问题环境的使用。

例 1 这是一个例子。

由例 1我们知道了例子环境的使用。

证明 1 这是一个证明。

由证明 1我们知道了证明环境的使用。

解 1 这是一个解。

由解 1我们知道了解环境的使用。

第六章 其它小功能

6.1 脚注

6.1.1 test

利用 \footnote{具体内容} 可以生成脚注[1]。

[1]脚注可以补充说明一些东西

6.2 无序列表与有序列表

无序列表是这样的：

- one
- two
- ...

有序列表是这样子的：

1. one
2. two
3. ...

6.3 字体加粗与斜体

如果想强调部分内容，可以使用加粗的手段来实现。加粗字体可以用 `\textbf{加粗}` 来实现。例如：**这是加粗的字体**。**This is bold fonts** 。

中文字体没有斜体设计，但是英文字体有。斜体 *Italics*。

第七章 参考文献与引用

参考文献对于一篇正式的论文来说是必不可少的，在建模中重要的参考文献当然应该列出。LaTeX 在这方面的功能也是十分强大的，下面进介绍一个比较简单的参考文献制作方法。有兴趣的可以学习 `bibtex` 或 `biblatex` 的使用。

LaTeX 的入门书籍可以看《LaTeX 入门》[1]。

参考文献

[1] 刘海洋. LaTeX 入门[J]. 电子工业出版社, 北京, 2013.

[2] 全国大学生数学建模竞赛论文格式规范 (2020 年 8 月 25 日修改).