浙江工业大学

本科毕业论文（设计）

文献综述

论文题目：Design and Implementation of Forbidden Word Recognition and Rendering with Browser Environment

学　　院: 计算机科学与技术学院

专　　业: 软件工程（中外合作办学）

班　　级: 2019 软件工程（中外合作办学）02

学　　号: 201906150218

学生姓名: 唐晨宇

指导老师: 李小薪

提交日期: 2023 年 03 月

Abstract: This paper is a literature review on the design and implementation of prohibited word recognition and rendering in a browser environment. The browser environment refers to the execution environment that parses JavaScript code with the Google V8 engine at its core, while prohibited word detection refers to the identification and processing of non-compliant content in content publishing platforms. In today's internet environment, the usage of internet media in enterprise publicity platforms is increasing. Under various legal constraints, it is necessary to ensure the compliance of content when publishing content, and a content compliance detection system is required for risk control. After considering the development principles of reusability, efficiency, and low cost in enterprise project development, the author proposes a solution for prohibited word recognition and rendering in a browser environment. Firstly, this paper introduces the origin of the project and its research significance. Secondly, this paper focuses on the analysis of existing file recognition technology, content security inspection technology, and browser rendering technology on the market, extracting the key points and difficulties of the project from both business and technical perspectives. Finally, based on the existing technology and analysis results, the advantages and disadvantages of various technology choices in system development are discussed, and the key technologies used are introduced.

Keywords: Content compliance, file recognition, PDF rendering, MS Office, Node.js, TypeScript

1 Introduction

With the continuous upgrading of modern financial enterprises' digital transformation, the number of users on online platforms has further increased, posing challenges to many of the enterprise's existing basic system architecture, such as the CRM platform, front-end marketing management platform, and information publishing platform[1], which is the focus of this paper. In 2016, General Secretary Xi Jinping proposed a series of "Internet+" plans and called for the establishment of a good network environment at the National Conference on Cybersecurity and Informatization, emphasizing that websites should bear the "main responsibility" in online information management and enhance the sense of mission and responsibility of internet enterprises to jointly promote the sustainable and healthy development of the internet[2].

Adhering to the concept of being a responsible enterprise, the author's financial enterprise has proposed the need for content compliance detection to enhance risk management in information publishing through machine auditing. Content compliance detection

is mainly manifested in the recognition of prohibited words in practical operation, with the ultimate goal of ensuring that the information published by the enterprise's content publishers on public platforms complies with laws and regulations, while also alerting the publishers to any problems in the content and enabling them to make timely corrections and publish content on schedule.

To accomplish the above tasks, it is necessary to start from a business perspective and analyze the required business processes from the content and platform of publication. In the overall process, the focus in front-end development is on how to speed up detection, display prohibited content in a reasonable manner, and thus improve the user experience. From a technical perspective, processing large files and displaying them in browser have become major challenges in completing the above tasks.

This article presents the following main contributions: (1) It elaborates on the research background and significance of content security in enterprise content publishing platforms after digital transformation; (2) It provides definitions of the concepts related to forbidden word recognition and rendering in the browser environment; (3) It analyzes the current workflow from a business perspective; (4) It summarizes the existing forbidden word recognition and rendering technologies based on the business process and points out that there is currently no technology that meets the current requirements; (5) It collects common front-end technologies; (6) It identifies the problems and challenges of the current project and discusses future research directions.

The organizational structure of this paper is as follows: Chapter 1 serves as an introduction, which mainly introduces the tasks and significance of forbidden word recognition and rendering in the browser environment, the research background, the main work of the article, and the organizational framework. Chapter 2 is the background and concept research, which mainly analyzes the research background of the project, provides relevant definitions, describes the research problems, and conducts business analysis. Chapter 3 discusses the current state of research, mainly introducing existing work in file recognition technology, content security inspection technology, browser rendering technology, as well as researching front-end development technology and frameworks that can be applied to this paper's work. Chapter 4 mainly discusses the problems and challenges faced in the work and future research directions. Chapter 5 is the conclusion, summarizing the work of this paper.

2 Research Background and Business Analysis

2.1 Research Background

With the widespread use of the Internet, more and more enterprises have started to undergo digital transformation by establishing their own digital platforms to drive business model transformation and publish information through emerging media platforms. In the financial enterprise, many banks in China have established fintech subsidiaries, while insurance and securities industries have also optimized their businesses through digital technologies, resulting in improvements in risk control and service levels[3].The successful integration of the internet and the financial enterprise has greatly improved the convenience and inclusiveness of financial services, but it has also brought significant risks. Given the complexity and specialized nature of financial products, ordinary consumers find it difficult to distinguish between genuine and fake products, and tend to rely on various media, especially influential ones, to obtain information[4], but these most authoritative media outlets are often owned by financial enterprise operators.

As socially responsible entities, particularly state-owned enterprises in the financial enterprise, it is crucial to prioritize the truthfulness, reliability, and effectiveness of financial information in media and advertising campaigns while also emphasizing risk management needs. To enhance compliance in media promotion, enterprises typically introduce review mechanisms and adopt the principle of position isolation and transparency, whereby content creators and reviewers are mutually independent. However, manual review is plagued by numerous issues, such as high workload, labor costs, and inconsistent standards. To effectively control costs and fulfill regulatory requirements, a growing number of enterprises have chosen to combine machine and manual review.

To ensure the compliance of content published on the platform, the content publisher first undergoes a preliminary check using machine learning algorithms to identify potentially negative keywords. If the content does not match any negative keywords, it will be automatically approved for publication. However, if the machine detects any potentially risky content, a manual review will be initiated. Any content that violates laws and regulations will be removed and blocked from the platform[5].The use of such auditing mechanisms is most common in user-generated content (UCG) platforms. For example, Mango TV has developed its own content security technology protection system, primarily targeting illegal and inappropriate content such as political, pornographic, and violent content in text, images, audio, and video formats. This is achieved through a

systematic integration of machine auditing, human auditing, user complaints, and results review[6], which providing valuable guidance for the establishment of system processes.

The importance of content moderation lies in its ability to help enterprises manage risks, reduce false information, prevent the spread of illegal content, protect the legitimate rights and interests of users, alleviate the pressures faced by risk control departments, and enhance the enterprise's sense of mission and responsibility. However, implementing content moderation is not easy and requires multiple technological approaches, including automated algorithms, deep learning, and natural language processing. Although there are many related research contents available online, there is still a significant gap in the actual project implementation. The fundamental reason is that different content platforms face different types and scales of content, and their specific implementation methods should be tailored to the specific business processes needed.

After considering all of the above, this article mainly analyzes the compliance issues faced by relevant enterprises in their promotion under the digital transformation of the financial enterprise. Based on actual development within the enterprise, the article verifies the feasibility of the project, delivers usable products quickly, and controls development costs. Of course, the business process proposed in this article is equally applicable to other content publishing platforms, such as government websites and school notice boards.

2.2 Business Analysis

When disseminating information online, we must pay attention to the following aspects: (1) Accuracy: ensure that the content posted online is accurate and true. If the content lacks authenticity, it will mislead customers, cause economic losses, and damage the enterprise‘s image; (2) Compliance: publishing content online may lead to legal issues, such as prohibiting the use of extreme terms and promises of returns in the financial enterprise, and avoiding guiding customers' product choices through implication;[7](3) Quality: the quality of content affects the user's reading experience, for instance, reducing the occurrence of typos can help improve the quality of content.

Compared to the UCG platform content moderation discussed in the previous article, enterprise information publishing platforms have specialized content and diverse publishing formats, which require higher risk management standards. The existing content moderation techniques do not fully meet business needs. Therefore, targeted analysis of the moderated content is required, such as short comments and news that only contain text, formal notifications that include both text and attachments, and so on. In addition, publishing platforms have specific requirements, such as WeChat public accounts, mo-

bile apps, and enterprise websites. Different business processes should be adopted based on the different moderated content and publishing methods to provide a better user experience. In addition to ensuring that non-compliant content is not published, we should also promptly identify specific issues found during moderation to assist content publishers in making targeted modifications.

The author proposes three core discussion points for this article: (1) How to obtain the content to be published (including text and files); (2) How to detect the published content for prohibited content; (3) How to display the prohibited words on the front-end page. In addition, we should also consider the following issues: (1) How to improve detection efficiency, making the detection process imperceptible and improving user experience; (2) How to reduce development costs and development risks in practical development; (3) How to ensure high cohesion and low coupling of the system, making it have good scalability.

3 Current Research Situation

According to the business process analyzed in chapter 2, we conducted relevant research on existing document recognition technology, content security inspection technology, and browser rendering technology in the market, analyzed and searched for available technologies, and pointed out the shortcomings of existing technologies, in order to find future research and improvement directions.

3.1 File Types and Recognition Techniques

In order to deal with the problem of attachments in formal notification, we need file recognition technology to determine the file type, so as to distinguish between different types of documents. At the same time, identifying file types can limit users to upload files that can not be processed by the system, reduce the possibility of system attacks, and improve the security and robustness of the system. In the detection system, we mainly receive MS Office file type and PDF file type, so in this section we mainly analyze these two file types.

3.1.1 File Header(Magic Number)

The Windows file manager identifies and determines the file type according to the suffix name of the file, that is, the last "." and the back part. However, the file name (including the suffix name) is completely transparent to the user and can be modified at will, and it is easy to encounter situations where the file name is tampered with or damaged, the file suffix name is missing or cannot be automatically recognized by the system, so relying on the extension name alone is not reliable[8]. In addition to the suffix name,

there is another way of file recognition-feature identification, that is, “Magic Number”, which refers to a string of character data stored in the file header to distinguish different file types. File feature identifiers can be identified by a string of hexadecimal strings, but do not have a fixed offset, that is, the number of hexadecimal bits varies from file to file. Except TXT, XML and other text type files do not have magic number, all the basic file types we use have this characteristic identification.

MIME type (Multipurpose Internet Mail Extensions type)[9] is a standardized description of file types, browsers can use MIME types to determine file types, Its default value is “text/plain”. There are usually three ways to detect MIME types[10]: (1) check the content-type carried in the request[11];(2) read the characteristic identification of the file; (3) judge according to the file name (suffix). In the browser environment, when a user uploads a file on a web page, the browser will use a series of file type identification algorithms, for example, the browser will read the byte stream of the uploaded file and parse the file feature identification information to determine the file type, and deal with it accordingly.

However, there are some problems in MIME type, for example, when the file characteristics of the file can not be accurately identified (the file features are not one-to-one corresponding to the type), the way of directly checking the suffix is adopted, which is an uncontrolled process in the program, so taking over the file identification module is of great significance to improve the security and expansibility of the system. For this reason, we need to further confirm the operation of the MIME type in the environment involved in this article through follow-up practice, and adjust accordingly according to the situation of the experiment.

3.1.2 MS Office

MS Office is a suite of applications developed by Microsoft that includes a range of productivity applications, including word processing, spreadsheets, presentations, database software, and so on. Among them, the documents, forms and presentations used in the office have come into the lives of almost all people who use electronic office. MS Office uses file formats with .doc, .xls, .ppt and other extensions, which is a composite document format. After 2007, a new version called Office Open XML (OOXML) was introduced[12], which usually refers to the file format with .docx, .xlsx, .pptx and other extensions[13]. The OOXML file format is based on the XML and ZIP file formats, where XML is an extensible markup language that can be used to identify data while allowing developers to add custom tags and extensions as needed, by adding metadata to the document. For example, XML emphasizes specific sentences in a report by using italics or

bold marks[14]. In addition to the extensibility of its basic format, OOXML also includes a range of tools and API to help developers read, create, and modify OOXML documents. For example, Microsoft provides Open XML SDK[15]. These features mean that they have better versatility and can be integrated with other software and platforms for data exchange and sharing. The OOXML file format also provides better data compression, making files smaller and faster to load. Due to the nature of ZIP file format in OOXML, it is treated as a single file when opened by an application, so it is well compatible with past versions of compound document formats.[13].Here are some descriptions of common MS Office file formats:

1. .docx: The default format of the Word document. It uses XML format instead of binary to store elements such as text, images, tables, charts, and so on.

2. .xlsx: The default format of the Excel workbook. It uses XML format instead of binary to store worksheets, charts, macros, and so on.

3. .pptx: The default format for PowerPoint presentations. It uses XML format instead of binary to store slides, charts, text boxes, animations, and so on.

In the work involved in this paper, MS Office is an important research object, because the information contained in the office file is easy to have some non-compliance information, and it is also easy to be used by lawbreakers to hide the information. The analysis and extraction of the content of this file is one of the key work of this paper, and dealing with multi-version and multi-type MS Office files is also a major challenge in practice.

3.1.3 PDF

PDF (Portable Document Format) is developed by Adobe Systems Inc. An electronic document format developed to facilitate the sharing and exchange of documents, to maintain consistent display on different operating systems and hardware platforms, to solve the problem of heterogeneity in electronic document transmission, and to become the standard for document release and transmission[16,17]. In this article, we note that PDF format has the following advantages: (1) multiple platforms: PDF can be viewed and printed on different operating systems and hardware platforms without compatibility problems; (2) easy to convert: files in other formats can be converted to PDF files, including MS Office mentioned above. (3) easy to transfer: compared with HTML files with the same content, the capacity of PDF files is 1/5; (4) strong practicability: support full-text search, tagging, comments and other functions; (5) high security: support

a variety of security features, such as password protection, digital signature and access control, which can protect the confidentiality and integrity of documents[18].

In addition, the PDF file format also has a series of advantages, such as open source, versatility, good compatibility and so on. Because of these advantages, PDF, as one of the important file formats in Internet transmission, has been taken into consideration. In 2023, most browsers will have built-in PDF viewers, and we can often quickly view the contents of PDF files without the need for additional software. Of course, if we want to do additional operations on PDF files in our project, we need more related technologies, such as rendering technology, which we will analyze in 3.3.1.

3.2 Content Security Review Technology

In the content security review technology at home and abroad, web page information evaluation mainly applies four filtering technologies, namely, filtering based on Internet content grading platform (PICS), database filtering (IP library, URL library), keyword filtering and intelligent content understanding filtering[19]. In this paper, keyword filtering and intelligent content understanding filtering are mainly used for content recognition, which mainly relies on the intelligent recognition technology of natural language processing technology NLP, deep learning and other technologies to achieve filtering[20]. Through the content analysis and understanding to generate a document model to identify prohibited content.

3.2.1 Natural Language Processing

NLP（Natural Language Processing）[21] is a branch of artificial intelligence which designed to enable computers to understand, interpret and generate human language. Natural language processing is a cross-discipline of linguistics, computer science and artificial intelligence, which studies the interaction between computer and human language, especially how to program computer to process and analyze a large amount of natural language data. The goal is to generate a model that can “understand” the content of the document, and this technology can accurately extract the information and opinions contained in the document, and classify and organize the document itself. This means that the computer must be able to recognize the grammatical and semantic structures in the language, analyze the words, phrases and sentences in the text, and correctly understand their meaning. When it comes to content security, NLP technology can be used in the following aspects: (1) Sensitive word filtering: NLP technology can be used to identify and filter sensitive words from text. It can recognize these words through rule-based methods or machine learning algorithms. For example, the content audit system uses

NLP technology to detect keywords such as violence, pornography, politics, terrorism and so on; (2) Text classification: NLP technology can be used to classify text into different categories. For example, the content audit system can divide the text into normal, illegal, and questionable categories, so that people can have a clear understanding of the audit objectives and make it easier to review. (3) Semantic analysis: NLP technology can be used to identify entities and concepts in text and understand the relationship between them. Because sometimes simple sensitive words do not violate the rules, only through semantics can we really find out the sensitive parts of the text.

Of course, rebuilding a content security audit model is not the focus of this article, but it is a necessary process to understand NLP. In a non-professional artificial intelligence research and development enterprise, we prefer to purchase related services for development rather than build our own natural language processing model from scratch. So this paper focuses on the applicability of these models in actual projects. Of course, we should also retain the possibility that it can be replaced.

3.2.2 Ali Cloud Content Security Service

Aliyun content Security Service[22,23] is an AI technology-driven content review service, mainly used for content security on Internet platforms. Content security products provide multimedia content risk detection capabilities such as pictures, video, voice, text to help users find risky content or elements such as pornography, violence and politics, which can greatly reduce manual audit costs and improve content quality, improve platform order and user experience. Its core technologies include multimodal computing, deep learning, natural language processing, image recognition and other AI technologies, which can achieve efficient and accurate content review. Aliyun content security service provides a variety of API invocation methods, including API and SDK. Users can choose different API invocation methods according to their actual needs. In addition, Aliyun content security service also provides a variety of audit policies and rules, which users can configure and adjust according to their business needs to meet different audit needs. Its main services are: (1) text security detection; (2) picture security detection; (3) video security detection; (4) voice security detection; (5) custom security policy. The service classifies the results of the detected content and returns the sensitive content in the indicated content. Aliyun content security service has been widely used in many industries to help users achieve fast, accurate and intelligent content review, improve the content security level of the platform, reduce the compliance risk of the platform, and protect the rights and interests and security of users.

Considering the development cost and the perfection of Aliyun service, we con-

sider using Aliyun content security service as an important tool for content audit in this system. However, because it is developed in a browser environment, Aliyun does not provide relevant SDK, so it is necessary to call API according to the corresponding rules.

3.3 Browser Rendering Technology

3.3.1 PDF.js

Usually browsers can open PDF files without plug-ins, but in the project we need to embed PDF pages to display the recognition results of prohibited words so that content publishers can modify them in a timely manner, so we need to find some pages that can display PDF on the front page. PDF.js is an open source JavaScript library developed by Mozilla for rendering PDF documents on Web, which can be used directly in Web browsers[24]. PDF.js uses HTML5 Canvas and CSS technology to render PDF documents. When a user requests to view an PDF document, PDF.js downloads and parses the PDF file asynchronously, and then renders it as an image on the HTML5 Canvas. In this way, users can view PDF documents directly in a Web browser without using a separate PDF reader. PDF.js provides a viewer.js for simple pdf to browser display, through the nesting of iframe can be collected in the system PDF browsing. PDF.js also has a flexible API that allows developers to customize their functionality as needed. For example, developers can use PDF.js API to extract metadata from PDF documents or to embed PDF documents in Web pages.

3.4 Front-end Development Technology and Framework

3.4.1 TypeScript

JavaScript is a relatively poor language for developing and maintaining large applications, and there are some difficulties in using JavaScript in a complex code base: language abstractions that lack robustness, such as static types, classes, and interfaces, which can hinder programmers' productivity and undermine tool support[25]. TypeScript is an extension of JavaScript designed to address this flaw, developed by Microsoft and a superset of JavaScript, so every JavaScript program is a TypeScript program[26,27]. TypeScript adds optional static types and some new features, such as classes, interfaces and so on. So that developers can do type checking and better maintainability when writing code. TypeScript is designed to provide lightweight help to programmers, so the module system and type system are flexible and easy to use. At the same time, it can also adapt to existing JavaScript projects very well, without the need for overall project rewriting[27]. Compared with JavaScript, TypeScript has the following advantages:

1. Static types: TypeScript can help developers detect type errors when writing code, reduce exceptions when the program is running, and improve the maintainability and readability of the code.

2. Classes and interfaces: TypeScript supports the definition of classes and interfaces, allowing developers to organize code and abstract data structures like compiled languages such as Java.

3. Ecosystem: TypeScript is widely used in Node.js, React, VUE and other large frameworks and libraries, which can improve the development efficiency of developers.

3.4.2 Node

Node.js[28] is a JavaScript running environment based on Chrome V8 engine, which has advantages in development efficiency and performance, and is widely used in the field of Web development. Node.js provides some special modules, which make it easy for developers to carry out I/O operation, process management, timer, database access and other operations, making it more convenient and quick to use JavaScript for server-side development. In addition to serving as a server, it is more used to quickly build extensible Web applications. For example, npm (Node Package Manager)[29] package management system is now the most important tool in front-end work, and it is also the default package management module of Node.js. With the widespread use of Node.js in Web application development, the Node.js community is also very active. Npm provides a large number of third-party modules and tools to help developers develop and deploy applications more quickly. More and more developers download npm packages to the local software environment, and develop software on this basis[30].

3.4.3 Promise

Node.js uses the single-threaded, event-driven, non-blocking I/O model, which not only brings a huge performance improvement, but also reduces the complexity of multithreaded programming, thus improving the development efficiency and making it lightweight and efficient[31]. When dealing with asynchronous problems, traditional Node.js generally uses callback functions. The callback function has a serious problem which called callback hell, that is, it is nested many times when the function runs in the order we want, which greatly affects the readability of the code. Promise is a solution to asynchronous programming, and Promise represents an operation that is not yet completed but is expected to be completed in the future; the Promise object represents the final

completion (or failure) of an asynchronous operation and its resulting value[32]. When using Promise, we can handle the results returned by the asynchronous operation by using the then () and catch () methods. Promise changes its state after the asynchronous operation completes, and it has three states: pending (waiting), fulfilled (successful), and rejected (failed); if the asynchronous operation is successful, the then () method receives a successful result, and if the asynchronous operation fails, the catch () method receives an error message.

4 Discussion

4.1 Problems and Challenges

In all the studies of the previous work in this paper, we can find that many technologies have been fully developed, but some of these technologies stay at the theoretical level and lack of practical project implementation. The ultimate goal of this paper is to select the technology type through these existing technologies, judge the advantages and disadvantages of the existing technology, select the appropriate technology for project development, and strive to create a compliant, efficient and scalable content publishing platform audit system. From the current research, the recognition and rendering of prohibited words in the browser environment mainly encountered the following problems and challenges:

1. Client performance: forbidden words recognition in browser environment needs to take into account the performance of client devices, including processor speed, memory size, network bandwidth and other factors. If you want to achieve real-time recognition, you need to ensure that the client can process the text quickly and efficiently and return the results;

2. Thesaurus establishment: the recognition of prohibited words in the browser environment needs to consider the problem of lexicon establishment. In order to achieve efficient and accurate recognition of prohibited words, it is necessary to build an appropriate thesaurus of prohibited words, and constantly update and maintain them according to the actual situation;

3. Browser restrictions: browser restrictions need to be taken into account in the recognition of prohibited words in the browser environment. Browser security policies may restrict access to certain resources, such as cross-domain requests and local file reads;

4. Multi-terminal adaptation: in practical applications, the system does not only run on browsers, but may need to be packaged as an application and run on multiple sides such as mobile.

5. Visual display of forbidden words: in addition to detecting the forbidden words of the published content, how to visually display the location of prohibited words for auditors is also a major challenge for us.

4.2 Future Research Direction

After raising the relevant questions, the author divides the future research direction into three main research directions according to the importance of business processes and issues:

1. File recognition is carried out in a controllable way, so as to reduce the amount of data transmission as much as possible and through text transfer rather than file transfer, so as to improve the real-time performance of the system;

2. Study the rules of interaction with Aliyun platform, and create an information transmission module that can be docked with the platform in the browser environment;

3. Study the method of rendering files on the browser side, so that users can intuitively have a certain understanding of the published content.

In addition to the above research direction, we can also carry out different designs and previews for different platforms, and develop from the perspective of multi-terminal operation, so that the system can adapt to a variety of environments.

5 Summary

This paper comprehensively investigates the research background and current situation of forbidden word recognition and rendering in the browser environment, and puts forward a detection scenario that meets the compliance requirements of the enterprise from the business point of view. Through the summary of previous work, this paper believes that an excellent forbidden word recognition and rendering system should have the following advantages: (1) It should cover the regular scenarios of content release in the enterprise, and has a certain expansibility to support different workflows; (2) It can complete detection without the perception of users; (3) It should control the amount of network data transmission and improve the efficiency of network transmission. (4) It

can process documents in time and has a timely feedback mechanism; (5) It is completed with as little development cost as possible; (6) It has a perfect responsibility system, and it should be responsible to editors and auditors. Finally, this paper focuses on the future challenges and opportunities, and points out that the biggest challenge in this field is to propose a complete workflow, verify its feasibility, and retain its scalability.

Bibliography

[1] 张倩. 现代金融企业数字化转型升级中存在的问题与解决对策 [J]. 文化创新比较研究, 2021(67-70).

[2] 陈力丹. 习近平在网络安全和信息化工作座谈会上的讲话 [J]. 新闻前哨, 2018(59-60).

[3] 谭西梅. 金融行业数字化转型的现状、挑战与建议 [J]. 商业文化, 2022(94-95).

[4] 张继红. 我国互联网金融广告行为的法律规制 [J]. 收藏, 2019, 5.

[5] 朱垚颖, 谢新洲, 张静怡. 安全与发展: 网络内容审核标准体系的价值取向 [J/OL]. 新闻爱好者, 2022(27-33). http://dx.doi.org/10.16017/j.cnki.xwahz.2022.11.008.

[6] 卢海波, 骆迅, 唐晔, et al. AI 赋能 + 合规导向 + 系统闭环：芒果 TV 内容安全保障技术体系的构建 [J/OL]. 广播电视信息, 2022(21-24). http://dx.doi.org/10.16045/j.cnki.rti.2022.06.022.

[7] 中华人民共和国广告法 [EB]. 2018.

[8] 张润峰. 基于特征标识的文件类型识别与匹配 [J]. 计算机安全, 2011(6): 40－42.

[9] MOZILLA. MIME types[EB/OL]. 2022. https://developer.mozilla.org/en-US/docs/Web/HTTP/Basics_of_HTTP/MIME_types.

[10] 轩蜗. Go 语言 MIME 类型介绍 [EB/OL]. 2020. https://xuanwo.io/2020/04-go-mime-intro/.

[11] FIELDING R, RESCHKE J. Hypertext transfer protocol (HTTP/1.1): Semantics and content[R]. 2014.

[12] van VUGT W. Open XML[J]. The markup explained, .

[13] PARK B, PARK J, LEE S. Data concealment and detection in Microsoft Office 2007 files[J]. Digital Investigation, 2009, 5(3-4): 104－114.

[14] HUNTER D, RAFTER J, FAWCETT J, et al. beginning XML[M]. [S.l.]: John Wiley & Sons, 2007.

[15] MICROSOFT. SDK 2.5[J]. URL: https://learn.microsoft.com/zh-cn/office/open-xml/open-xml-sdk (date of the application: 19.04. 2020), .

[16] BIENZ T, COHN R, ADOBE SYSTEMS (MOUNTAIN VIEW C. Portable document format reference manual[M]. [S.l.]: Citeseer, 1993.

[17] CASTIGLIONE A, DE SANTIS A, SORIENTE C. Security and privacy issues in the Portable Document Format[J]. Journal of Systems and Software, 2010, 83(10) : 1813 – 1822.

[18] 陈婷玉, 苑丽华. PDF 格式在成果地质资料汇交工作中的优点 [J]. 档案, 2009(50-51).

[19] 彭昱忠, 元昌安, 王艳, et al. 基于内容理解的不良信息过滤技术研究 [J]. 计算机应用研究, 2009(433-438+447).

[20] 王宏宇, 陈冬梅. 网络信息内容安全技术浅析 [J/OL]. 电脑知识与技术, 2018(51-52).
http://dx.doi.org/10.14004/j.cnki.ckt.2018.0492.

[21] Wikipedia contributors. Natural Language Processing[EB]. 2023.

[22] 阿里云. 阿里云内容安全文档 [EB]. .

[23] 张建军, 孙滔, 孟方. 通过人工智能实现内容智能审核及在世界杯的实战 [J]. 现代电视技术, 2018(52-54+145).

[24] Mozilla. Getting Started with pdf.js[EB]. .

[25] RASTOGI A, SWAMY N, FOURNET C, et al. Safe & efficient gradual typing for TypeScript[C] // Proceedings of the 42Nd Annual ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages. 2015 : 167 – 180.

[26] BIERMAN G, ABADI M, TORGERSEN M. Understanding typescript[C] // ECOOP 2014–Object-Oriented Programming: 28th European Conference, Uppsala, Sweden, July 28–August 1, 2014. Proceedings 28. 2014 : 257 – 281.

[27] FENTON S, FENTON, SPEARING. Pro TypeScript[M]. [S.l.] : Springer, 2014.

[28] Node.js[EB]. 2022.

[29] npm – build amazing things[EB]. 2022.

[30] GOSWAMI P, GUPTA S, LI Z, et al. Investigating the reproducibility of npm packages[C] // 2020 IEEE International Conference on Software Maintenance and Evolution (ICSME). 2020 : 677 – 681.

[31] 邓森泉, 杨海波. Promise 方式实现 Node.js 应用的实践 [J/OL]. 计算机系统应用, 2017(218-223).
http://dx.doi.org/10.15888/j.cnki.csa.005700.

[32] MOZILLA. Promise[EB]. n.d..