浙江工业大学

本科毕业论文（设计）

开题报告

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

Design and Implementation of Forbidden Word Recognition and Rendering with Browser Environment

1 Background and Significance

1.1 Purpose and Significance

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.

Starting from the enterprise "compliance content detection", this paper studies the detection and rendering of prohibited words in the browser environment, in order to promote the construction of enterprise content compliance. Today, with the rapid develop-

ment of the Internet, the number of users of the online platform is also increasing, the risk management needs of enterprises are also constantly improving, and the content compliance testing system has a lot of market space and social demand.

The significance of the project research is to effectively reduce the compliance risk of content published on the platform and prevent the occurrence of illegal events caused by improper content release. The research of forbidden word detection and rendering system is to enable enterprises to strengthen the risk management of enterprise media content in a low-cost way, and is an effective means to reduce the risk of enterprise content release. It is conducive to the sustained and healthy development of the Internet environment. At the same time, we also focus on open source technology, through modular development to inject vitality into the open source community, to provide help for more newcomers.

From the perspective of digital transformation of the financial enterprise, this paper aims to solve a series of problems in the audit process of enterprise content platform, and is also applicable to other content publishing platforms, such as government websites, school notice release and so on. Also want to make a modest contribution to the good development of the Internet.

1.2 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 trans-

parency, 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.

1.3 Present Situation of Research

1.3.1 Content Review Platform

Content moderation mechanisms are commonly found on UGC platforms, such as Mango TV, which has its own proprietary content safety technology system. However, not all enterprises develop their own content safety review systems. To reduce review costs, improve review quality and efficiency, many third-party content review agencies have entered the content moderation outsourcing service enterprise, mainly including traditional mainstream media, internet cloud service giants, AI technology enterprises, and specialized content moderation outsourcing enterprises[7].

However, whether it is a self-developed platform or a third-party service, the audit process of "machine audit + manual audit" can generally be summarized as "machine marking classification-machine preliminary examination (AI marking)-manual review-manual third trial-manual sampling inspection" and so on[7]:

1. Machine marking classification：Make use of the content analysis ability of AI, tag the content, store it in different repositories, and enter the preliminary examination;

2. Machine preliminary examination： AI will match the contents of the library to the relevant check libraries, such as "sensitive word library" and "sensitive picture library", and mark the possible problems, so as to provide reference for the follow-up review;

3. Manual review: Judge the tag and classification label of the first trial, and read through the work in a comprehensive way, and give a judgment on whether the content is in compliance.

4. Manual third trial: It is mainly to confirm and deal with the results of the review, and to be responsible for making decisions on the contents of the review;

5. Manual sampling inspection: The content works that have been examined three times need to be sampled manually at last.

In the system designed in this article, the goal is to design a content audit system that meets the requirements of the enterprise, ensuring that content is audited as quickly and efficiently as possible through a fast and effective feedback mechanism. The details of machine content audit are described in more detail in 1.3.3.

1.3.2 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.

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. At present, there are indeed some developed MIME type recognition packages, but after research, it is found that it does not support the MS Office format we need[12], which uses the open file format of Office Open XML (OOXML)[13]. Its content identification is consistent with the compression type, so additional development is needed to complete this work.

1.3.3 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[14]. 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[15]. Through the content analysis and understanding to generate a document model to identify prohibited content.

NLP（Natural Language Processing）[16] 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.

Aliyun content Security Service[17,18] 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 consider 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.

1.3.4 Browser Rendering Technology

1.3.4.1 PDF.js

Usually browsers can open PDF[19,20] 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[21]. 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.

In the existing projects of the enterprise, we mainly use pdf.js for pdf browsing, but the file reading mode and how to output after the forbidden words are identified are not given in the viewer.js template, so we need to carry out secondary development to achieve the function of text reading and rendering.

2 Contents and Objectives

2.1 Research Objectives

Based on the business requirements presented by the enterprise, this article compared and analyzed various content security audit systems, and identified a range of possible technologies that can be adopted to meet the needs of the project. In the current environment of continuous digital transformation of modern financial enterprises, it is crucial for internet enterprises to enhance their sense of mission and responsibility, and promote the sustained and healthy development of the internet. Leveraging information technology to improve service quality has become an important means for enterprise development. Internet enterprises have already achieved successful development in this regard, and traditional enterprises should also continue to explore the possibilities of digital transformation in their businesses to improve service quality in the internet era. In order to enhance enterprise compliance and strengthen the application of digital intelligence

in business, major platforms have made their own explorations and developed digital departments. The research objective of this project is to use JavaScript-related technologies to implement a system for identifying and rendering prohibited words in the browser environment, especially by applying some cutting-edge technologies and frameworks in JavaScript, such as TypeScript, Node, and Vue.

2.2 Basic Contents

In a non-artificial intelligence specialized enterprise, we do not need a content recognition model, but rather focus on business implementation. In this study, we mainly focus on the design and implementation of prohibited word recognition in the browser environment. Its main function is to perform real-time prohibited word detection when content publishers post notices on the platform, and to promptly return the detection results, with a clear display of the results. The Figure 1 shows the process flow of the Program.

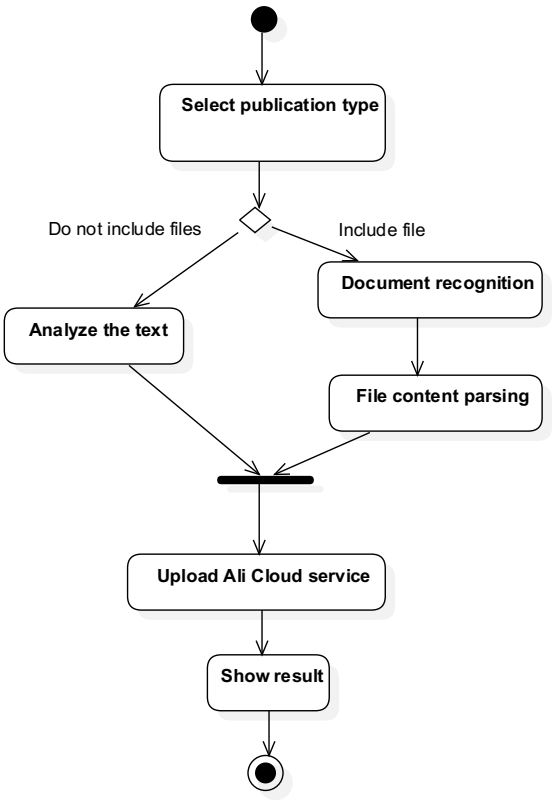

Figure 1: Activity diagram

The specific contents of this study include:

(1) File type identification

The main function of file type identification is to identify the file types uploaded

by users to facilitate subsequent processing. The main file types are PDF and MS Office documents. When the authorized content publisher has the need for information release, if the content has the need for file upload, the user will identify the type of the file when uploading the file, and issue a corresponding warning to the non-compliant file type.

(2) Text extraction of files

The text extraction is an operation performed after identifying the file type. To reduce data transmission and solve the issue of size limitation in Aliyun file detection, the system extracts content based on the identified file type. By parsing binary files, the system converts file content into a format suitable for network transmission.

(3) Format conversion of files

In case of problematic files, we need to display them on the web page. Therefore, we may need to convert such files to a format that can be easily displayed on the browser. PDF is a well-suited document format for sharing and exchanging, which maintains consistent display effects across different operating systems and hardware platforms. In practice, if any prohibited content is found, the file should be converted to PDF format and displayed on the browser.

(4) Review of content

The content audit mainly relies on calling the API of Ali Cloud Content Security Service. In this study, there are two research directions: API calls through the server or directly in the browser. In subsequent research, the appropriate direction will be selected for development based on security, efficiency, and other considerations. However, both solutions need to meet the Alibaba Cloud API calling rules and go through a series of content encryption methods to work properly.

(5) The display of forbidden words

If a prohibited word is detected in the published content, we will highlight the violation in the preview interface for text content. For file content, we will use an embedded page to display the content, and use a cross-referencing method to help users quickly locate the position of the prohibited word, making it easy to make modifications in the original file.

In addition to the functional requirements mentioned above, we should also consider the following system characteristics: (1) Scalability: Ensure that the system has good expandability; (2) Security: Ensure that information is not leaked during transmission; (3) Portability: The system can be deployed in existing platforms and can be transferred to different projects; (4) Maintainability: With easily understandable code that can be maintained by other developers; (5) Efficiency: With low time delay; (6) robustness:

With good risk resistance when handling files, so that the system will not crash during file processing.

2.3 Technical Difficulties

1. Implementing file recognition in a controllable manner and reducing data transmission volume by using text transfer instead of file transfer as much as possible to improve the real-time performance of the system;

2. Extract the content of different versions and different types of files to solve the problem of Ali cloud file detection size limit (5M), which means that we need to manipulate the binary files in the browser environment;

3. 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;

4. Study the method of rendering files on the browser side, so that users can intuitively have a certain understanding of the published content. It mainly through the research on the source code of pdf.js, through exploring its working logic to modify the file acquisition and content highlighting code to meet the system requirements;

5. Cross-page object acquisition and operation of Iframe;

6. Carry out multi-terminal adaptation, so that a system can run on multiple platforms, meet the different needs of different services, and improve the availability of the system.

2.4 Expected Research Objectives

The expected research objectives of this topic are as follows:

1. Complete the implementation of the forbidden word recognition and rendering system in the browser environment. Users only need to input the content and files to be published, and the system can automatically extract the required data and upload it to the cloud service, and feedback the corresponding results in time;

2. The modular design of the system packages and encapsulates its functions, providing the open-source community with the implementation of a series of functions such as file recognition and Alibaba Cloud service requests;

3. Strive for multi-end adaptation, so that the system can meet a variety of needs of content release, improve the user experience;

4. Let the results of the project be applied in the real environment. At present, the related system is piloted in a financial enterprise, so that the smooth industrialization of scientific and technological achievements of research and development is also one of the goals of this task.

3 Technical of Research

(1) System platform: macos Ventura

(2) System architecture: B/S architecture

The B/S (Browser/Server) architecture is a software application architecture based on Web browsers and servers. In the Web architecture, the user sends a request to the server through the Web browser, and the server responds to the request and returns the web page to the user browser, realizing the interaction and data transmission of the application program. With the rise of Internet technology, it is a kind of structure that changes or improves the structure of C/S. Compared with C/S structure, B/S architecture has the advantages of cross-platform, portability, easy maintenance and management. In this topic, we mainly carry out the development and testing task through the Chrome browser.

(3) Programing language: 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[22]. 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[23,24]. 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[24].

(4) Architecture used: Vue+Node

Vue.js is a popular JavaScript front-end framework for building interactive Web interfaces. By combining templates and data models, adopting two-way data binding, component development, virtual DOM and other ways. Developers can easily create a

reusable and maintainable user interface, and carry out good data management and efficient page rendering.

Node.js[25] 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)[26] 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[27].

(5) Server software: Nginx

Nginx is a high-performance open source Web server software that can also be used as a reverse proxy, load balancer, HTTP cache, and security controller. It uses event-driven asynchronous architecture, can handle thousands of concurrent connections, and has the advantages of low latency and high throughput. Nginx also has extensibility and a high degree of customization, and can achieve different functions through many third-party modules.

(6)System development tools: Visual Studio Code

Visual Studio Code (VS Code for short) is an open source, cross-platform code editor developed by Microsoft. It supports multiple programming languages and frameworks, and has a rich plug-in ecosystem to help users develop and debug code more efficiently. At present, there are many plug-ins supporting the front-end on VS Code, which can effectively improve the development efficiency, and it is also the preferred code editor for many front-end engineers.

(7) Research on other cutting-edge related technologies

In addition to the technologies used above, this paper will also explore the deployment of micro services, Docker and other technologies, as well as the availability of Electro technology packaged as desktop technology in Web applications, and explore the application of new technologies in practice.

4 Schedule

Table 1: Paper work arrangement

Task serial number	Between start and stop	The key points of the stage.
1	2022.12.27-2023.1.20	Understand the relevant contents of the topic and look up materials in Chinese and English
2	2023.1.30-2023.3.03	Consult the literature, complete the literature review, opening report and foreign language translation
3	2023.3.03-2023.3.20	Learn TypeScript programming and study the related applications of Vue and Node
4	2023.3.20-2023.3.27	Complete the basic page construction of the project
5	2023.3.27-2023.4.10	Complete file identification and parsing
6	2023.4.10-2023.4.17	Complete the API docking of Ali Cloud
7	2023.4.17-2023.4.24	Complete page rendering
8	2023.4.24-2023.5.1	Complete deployment testing
9	2023.5.1-2023.5.24	Collate the materials and complete the graduation thesis
10	2023.5.24-2023.6.18	Hand in graduation thesis and prepare graduation defense.

References

[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] 黄孝章, 童婷薇. 互联网第三方内容审核服务发展探析 [J/OL]. 北京印刷学院学报, 2023(50-56). http://dx.doi.org/10.19461/j.cnki.1004-8626.2023.01.011.

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

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

[14] 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).

[15] 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.

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

[17] 阿里云. 阿里云内容安全文档 [EB]. .

[18] 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).

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

[20] 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.

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

[22] 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.

[23] 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.

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

[25] Node.js[EB]. 2022.

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

[27] 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.