Design and Application of a Statistical Analysis System for Electronic Resource Usage Based on Fiddler Proxy Program: Postprint

Authors: Chen Guang

Date: 2023-08-27T00:00:00+00:00

Abstract

[Purpose/Significance] This study conducts a comparative analysis of the primary methods employed by domestic libraries for independently collecting electronic resource usage statistics, addressing the limitation of existing technical approaches in capturing HTTPS-based electronic resource access information.

[Method/Process] The technology was enhanced upon the existing bypass monitoring collection paradigm, implementing statistics and analysis of HTTPS-based electronic resource usage data through policy routing and Fiddler proxy programs. An electronic resource usage statistics and analysis system was subsequently designed and deployed.

[Results/Conclusion] The collection challenge for HTTPS-based electronic resource usage data was successfully resolved, offering valuable reference for other libraries undertaking independent collection of electronic resource usage statistics.

Full Text

Preamble

Vol. 62 No. 13, July 2018, ChinaXiv Cooperative Journal

Design and Application of Electronic Resource Usage Statistical Analysis System Based on Fiddler Proxy

Fujian Institute of Research on the Structure of Matter, Chinese Academy of Sciences, Fuzhou 350002

Abstract

[Purpose/Significance] This paper compares and analyzes the main methods currently used by domestic libraries to independently collect electronic resource

usage statistics, addressing the problem that existing technical methods cannot capture electronic resource access information based on the HTTPS protocol. [**Method/Process**] Building upon the existing bypass monitoring collection mode, the technology is improved by implementing policy routing and a Fiddler proxy program to achieve statistics and analysis of electronic resource usage data accessed via HTTPS protocol. Based on this approach, an electronic resource usage statistical analysis system is designed and applied. [**Result/Conclusion**] The system solves the data collection problem for electronic resource usage based on HTTPS protocol access, providing a valuable reference for other libraries seeking to independently collect their own electronic resource usage statistics.

Classification Number: G250.7
Keywords: electronic resources, policy routing, usage statistics data, Fiddler proxy, bypass monitoring
DOI: 10.13266/j.issn.0252-3116.2018.13.005

Electronic resources such as e-books, e-journals, full-text databases, and secondary abstract databases have gradually become the primary means for library users to access information. Their proportion in library collections continues to grow, with some libraries completely abandoning print resources and allocating their entire budget to electronic resource procurement. As electronic resources play an increasingly important role in libraries, usage statistics have become a critical tool for analyzing and understanding their value. These statistics accurately reflect utilization patterns and provide important references for libraries to restructure network portals, develop user training programs, highlight key electronic resource products, and assist librarians in making collection development decisions regarding electronic resource acquisition and management.

Database vendors' usage statistics reports are the primary means for libraries to obtain electronic resource usage data. Most vendors now provide statistics compliant with the COUNTER usage statistics and data measurement standards, supporting the SUSHI standard harvesting interface for unified collection and integration, enabling management of all electronic resource usage statistics. While COUNTER-compliant reports offer advantages in standardization, usability, automation, and low cost, they have certain limitations: (1) **Timeliness:** Vendor systems typically generate monthly reports around the middle of the following month, creating a time lag in data acquisition. (2) **Comprehensiveness:** COUNTER reports only cover subscribed resources, not the library's complete network resources (such as free electronic resources and institutional repository resources). (3) **Granularity:** COUNTER reports provide only statistical numbers, failing to meet libraries' needs for timely, in-depth content-level and user-level analysis and data mining.

To address these COUNTER limitations, domestic libraries have conducted research and practical applications on independently collecting electronic resource usage statistics. Based on existing research papers, these self-collection methods can be divided into two main categories: gateway log-based collection and analysis mode, and bypass monitoring-based collection and analysis mode.

2. Two Methods for Libraries to Collect Electronic Resource Usage Statistics

2.1 Gateway Log-Based Collection and Analysis Mode

All user Internet access data passes through unified exit gateways (such as core switches, firewalls, or proxy servers), which record all access information in logs, including electronic resource access data. By collecting, filtering, and analyzing these logs, libraries can generate electronic resource usage statistics reports. Domestic implementations include Yan Xiaodi et al.'s design of an electronic resource utilization statistics gateway, Wang Xiaoliang et al.'s construction of an e-journal database statistics analysis system through firewall log mining, Guo Zhenying et al.'s design of an electronic resource log statistics system, and Zhou Xin et al.'s analysis of user behavior through Web log mining.

Research on these cases reveals that the gateway log-based approach offers the advantage of not requiring additional network equipment for data capture, as it directly utilizes existing gateway logging functions. After technical processing to filter, clean, analyze, and integrate the data, electronic resource usage reports can be generated, saving hardware costs. The key technical challenge lies in the fact that gateway logs contain not only electronic resource access data but also other network access information. Efficient, rapid, and accurate matching of IP addresses and characteristic values is required to filter out irrelevant data.

However, this mode has limitations: (1) **Data Source Dependency:** Statistics reports rely solely on gateway log data, which may be incomplete and insufficient for in-depth analysis and data mining. Log formats vary significantly across different gateways, making relevant data processing code non-universal. (2) **Reporting Lag:** Report generation timing depends on log harvesting strategies. Daily harvesting enables daily reports, while monthly harvesting only allows monthly reports. Due to the nature of log data, real-time harvesting is impossible, making real-time monitoring of electronic resource access behavior unachievable.

2.2 Bypass Monitoring-Based Collection and Analysis Mode

The bypass monitoring mode operates by using port mirroring on network exit core devices to duplicate data flows, forwarding the copied data to a monitoring and analysis server. This server captures packets, parses access data, filters and analyzes information, generates usage statistics reports, and monitors user electronic resource access behavior throughout the process. Domestic implementations include Zhu Ling et al.'s evaluation of data acquisition quality in ERU and DRAS monitoring systems, Zhang Jilong et al.'s research on library user information behavior data collection methods based on the ERU system, Zou Rongli et al.'s design and application of electronic resource access management and control systems using bypass monitoring, Shi Xiaohua et al.'s design and application of university electronic resource access management and control systems, Wang Zhengjun et al.'s design and implementation of digital

resource evaluation systems based on bypass monitoring, and Wu Qunhui et al.'s research on university library electronic resource usage statistics models for scientific research.

The bypass monitoring mode offers several advantages: (1) It directly parses data packets from user access requests and electronic resource responses, obtaining the most original, accurate, and complete access information. (2) Packet mirroring, copying, and parsing occur in real-time with user information access behavior, enabling real-time monitoring and report generation. Timely data monitoring helps strengthen regulated use of electronic resources by setting violation thresholds, providing early warnings and handling violations to prevent database vendors from imposing large-scale bans that would affect other users' normal access. (3) As a bypass mode that copies packets without modifying original data or changing network topology, it has no impact on user network access behavior.

The key technical challenge in bypass monitoring lies in packet capture and parsing. When developing their own monitoring systems, libraries can use WinPcap or WireShark on Windows platforms, or the NetFilter framework or Iptables firewall on Linux platforms. For libraries lacking technical capacity, commercial software such as Shanghai Guanghua Fudan Company's ERU system or Tongfang CNKI's DRAS system can be purchased for monitoring and analysis.

Compared with gateway log-based mode, bypass monitoring offers stronger timeliness, more comprehensive and accurate data, and no impact on network structure or user behavior. The technology is mature, and commercial software is available, making it currently the most suitable mode for libraries to independently collect electronic resource usage statistics. Its only limitation is the need for additional dedicated monitoring servers, requiring extra hardware investment and higher costs.

3. Problems in Existing Modes and Solutions

3.1 Problems in Existing Modes

During testing of both modes, the author discovered that while HTTP protocol-based electronic resources could be captured by both methods, HTTPS protocol-based resources could not be captured by either gateway logs or monitoring servers. In other words, both modes are ineffective for HTTPS-based electronic resources.

Analysis reveals that the key difference between HTTP and HTTPS lies in their data transmission methods: HTTP uses plaintext transmission, while HTTPS uses encrypted transmission. HTTP-transmitted data is transparent to all network devices along the link, allowing gateways or monitoring servers to directly access and process the content when packets arrive. HTTPS encryption occurs at the application layer in the OSI seven-layer model, with encryption completed before packets are sent from the network adapter. When packets reach

gateways or monitoring servers, they are already encrypted, making it impossible to decrypt the ciphertext. Only source IP address, destination IP address, and domain name can be captured—insufficient for generating usage statistics reports or real-time monitoring.

Despite HTTPS's slower access speed and higher deployment costs, its superior encryption performance effectively prevents information leakage, prompting more websites to shift from HTTP to HTTPS. In electronic resources, foreign databases such as ScienceDirect, Nature, OSA, and Springer have already adopted HTTPS access. With server hardware development and technological advances, HTTPS deployment costs will decrease while access speeds improve significantly. It is foreseeable that more electronic resource providers will deploy HTTPS access for information security considerations. For libraries, addressing the challenge of capturing user information behavior from HTTPS-based electronic resources has become an urgent priority.

3.2 Technical Improvements to Bypass Monitoring Mode

Several methods can decrypt HTTPS traffic: (1) Using private keys stored in browsers to decrypt server-returned encrypted data; (2) Capturing required information before server data encryption or after decryption at the application layer; (3) Employing man-in-the-middle technology to control communication between client and server. The first two methods require client software or plugins on user PCs, significantly impacting users and making them unsuitable for libraries. The third method—man-in-the-middle technology—can create connections with both communication ends separately, exchanging received data so that both ends believe they are communicating directly while the entire session is actually controlled by the intermediary. Users cannot perceive the impact of this technology on their information access behavior. This characteristic makes man-in-the-middle technology most suitable for libraries to capture HTTPS-based user information behavior.

By deploying a man-in-the-middle technology-supporting program on the monitoring analysis server to replace the original packet capture program, HTTPS-based user information behavior capture can be achieved. Among such software, Fiddler is a powerful, free Web proxy program that can parse HTTPS protocol and record HTTP/HTTPS access information with simple configuration, making it the optimal choice for this study's intermediary program.

Unlike traditional bypass monitoring that uses port mirroring to copy packets, man-in-the-middle technology requires establishing connections with both users and electronic resource providers separately. This necessitates using policy routing (PRB) on core switches instead of port mirroring to forward user access packets to the intermediary program. Policy routing is a mechanism for route selection based on user-defined strategies that can forward data packets through configuration of access control lists (ACLs) containing electronic resource server IP addresses, enabling packet forwarding and filtering functions.

In summary, this study improves upon the existing bypass monitoring collection mode by using policy routing instead of port mirroring for packet filtering and forwarding, and deploying a man-in-the-middle technology-supporting program on the monitoring analysis server instead of a standard packet capture program. This enables collection of HTTPS-based user information behavior and forms the foundation for designing and applying the electronic resource usage statistical analysis system.

4. System Function Module Design and Implementation

4.1 System Function Module Design and Business Process

4.1.1 System Function Module Design The electronic resource usage statistical analysis system consists of four functional modules: data filtering and forwarding module, data analysis module, statistical analysis module, and violation monitoring module, as shown in [Figure 1: see original paper].

The data filtering and forwarding module filters packets on core switches and forwards electronic resource-related packets to the monitoring analysis server. The data analysis module deploys a Fiddler proxy program on the monitoring analysis server to receive and analyze user electronic resource access data, matching user full-text access behaviors based on URL characteristic values and HTTP status codes, then records them in an SQL database. The statistical analysis module uses a B/S architecture with specific SQL statements to display user electronic resource access information. The violation monitoring module uses C# programs to monitor user full-text access frequency, taking warning and banning actions against violators.

4.1.2 System Business Process When users generate Internet data, the data filtering and forwarding module filters out non-electronic-resource-related data and forwards electronic resource access data to the data analysis module. The data analysis module analyzes the data, records full-text access behaviors from user electronic resource access data, and simultaneously triggers the violation monitoring module. The violation monitoring module updates corresponding users' full-text download counts, compares them with user-preset daily warning and disabling thresholds, and takes corresponding measures based on comparison results. The statistical analysis module operates independently from this business process, providing visualization of electronic resource access data.

4.2 System Function Implementation

4.2.1 Data Filtering and Forwarding Module The data filtering and forwarding function is implemented by enabling policy routing on core switches and configuring corresponding ACLs containing electronic resource server IP addresses. The filtering function matches packets based on destination IP addresses—when packets arrive at the core switch, their destination IP is extracted and matched against the ACL list. Packets matching electronic resource

server IPs are forwarded to the monitoring analysis server (192.168.4.45), while non-matching packets are left unprocessed.

Using Wiley database as an example, the ACL matching code is as follows:

```
acl number 3200 description wiley
rule 1 permit ip destination 199.171.202.195  // Wiley database server IP
if-match acl 3200
apply ip-address next-hop 192.168.4.45  // Monitoring analysis server IP
```

Some electronic resource servers use CDN (content delivery network) acceleration with frequently changing IP addresses, requiring dynamic ACL list maintenance to prevent packet loss and missing user access data. A C# program tracks current IP addresses of each electronic resource server every 10 minutes, comparing them with ACL list IPs. If the ACL list doesn't contain a particular IP, the program updates the ACL list through the core switch interface.

4.2.2 Data Analysis Module The data analysis module deploys the Fiddler proxy program on the monitoring analysis server to capture and analyze user electronic resource access data. Through characteristic value matching and conditional judgments, it records qualifying user information behaviors in an SQL database.

When Fiddler receives a user access request, it first extracts the target server domain information to determine the corresponding URL characteristic value. It then matches the user request URL against the characteristic value. For requests satisfying the characteristic value, it checks whether the HTTP status code is 200 to confirm a successful request. Next, it verifies that the data returned from the electronic resource server is not empty. Finally, it marks requests meeting all these criteria and records relevant access information in the SQL database. The business process is shown in [Figure 2: see original paper].

The SQL database creates separate databases for different electronic resources, using "T+year" as table names to store annual user full-text access information. Tables primarily contain username, department/research group, user IP, access URL, time, and other information. Detailed field settings are shown in .

Table 1: Full-Text Access Data Table

Field	Type	Description
id	nvarchar(1000)	Unique identifier field
url	nvarchar(1000)	User-accessed full-text URL
ip	nvarchar(20)	User IP address
username	nvarchar(100)	Username
year	char(4)	Year
month	char(2)	Month
day	char(2)	Day
hour	char(2)	Hour

Field	Type	Description
minute	char(2)	Minute
second	char(2)	Second
researchgroup	nvarchar(30)	User's research group or department
type	nvarchar(10)	Full-text type accessed by user (PDF or HTML file)

4.2.3 Statistical Analysis Module The statistical analysis module uses a B/S architecture with PHP and SQL statements to create web pages displaying electronic resource usage statistics, utilizing HighCharts plugins for data visualization. The module provides different page content for libraries and research groups.

The library page displays full-text download counts and proportions for each electronic resource using pie charts and tables, while also providing individual resource research group total download counts and corresponding cost information (see [Figure 3: see original paper]). The research group user page similarly displays download counts and costs using pie charts, bar charts, and tables, and provides a warning settings page for configuring daily warning and disabling thresholds to prevent excessive downloading (see [Figure 4: see original paper]).

4.2.4 Violation Monitoring Module The violation monitoring module uses a C# program to monitor the full-text access data table in the data analysis module. The C# program records daily full-text download counts for each research group in real-time. When a group's daily downloads reach preset warning or disabling thresholds, corresponding actions are triggered.

The C# program's sqlDependency class provides a function that automatically triggers an OnChange event to notify the application when monitored data tables change. The program uses the sqlDependency class to monitor the full-text access data table—when new data is written, the OnChange event is triggered. The program extracts the id, researchgroup, and ip fields from the newly added record, increments the corresponding group's full-text download count by 1, and then evaluates the updated count. When the count reaches the daily warning threshold, the program sends a warning email to designated addresses; when reaching the disabling threshold, it immediately suspends the group's database access rights while sending a notification email. The program automatically resets all groups' download counts to zero at midnight daily. The program interface is shown in [Figure 5: see original paper].

5. System Application Effects and Existing Problems

5.1 System Application Effects

The electronic resource usage statistical analysis system was deployed in July 2017 and has been running stably after multiple modifications. As of January 20,

2018, the system has stored over 360,000 user full-text access records, providing powerful data support for the library.

Regarding electronic resources, the statistical analysis module enables understanding of each resource's full-text download volume, proportion, and cost per article, helping libraries grasp usage patterns and adjust resource guarantee strategies. For example, the OSA database had low full-text downloads in 2017, accounting for only 0.6% of total downloads with the highest cost per article (¥54.18) among all library resources. Analysis of full-text URLs revealed that most downloads came from the database's free OA journals. Based on this information, the library could consider discontinuing the subscription and providing access through alternative means.

Regarding user information behavior analysis, the statistical analysis module reveals research group preferences for electronic resources, helping libraries understand research directions and needs to better develop knowledge services. Additionally, precise full-text download data enables accurate calculation of cost-sharing proportions for charging research groups partial electronic resource fees.

Regarding violation monitoring, the system detected abnormal full-text download volumes from a research group in the ACS database in September 2017. Analysis revealed the group was using EndNote software for batch downloading. The system's timely warning enabled administrators to temporarily suspend the group's ACS access rights, preventing database vendor bans that would have affected other users.

5.2 Existing Problems in the System

The system uses policy routing instead of port mirroring, directly forwarding user access data to the monitoring analysis server. If the monitoring server fails and cannot forward user access data to electronic resource servers, users cannot access resources. This could be addressed using a backup monitoring server that activates when the primary server fails, ensuring resource access, though this introduces additional hardware costs.

Some electronic resource server IP addresses change frequently. Although dynamic IP list updates are implemented with a 10-minute interval, when server IPs change and the ACL list hasn't been updated yet, user access data cannot be forwarded to the monitoring analysis server, resulting in missing usage statistics. This could be resolved through access control that only permits electronic resource access from packets forwarded by the monitoring analysis server, forcing all user access through the server to ensure complete statistics. However, this approach would cause temporary access failures when resource server IPs change before ACL updates.

References

[1] Chen Daqing, Ye Lan, Yang Wei, et al. Design and implementation of electronic resource usage statistics platform USSER[J]. Library and Information Service, 2015, 59(1): 106-112. [2] Cai Jing. Research on statistics and standardization of electronic journal usage data[J]. Digital Library Forum, 2012(7): 64-69. [3] Zhu Ling, Cui Haiyuan. Discussion on evaluation methods for data acquisition quality of electronic resource monitoring and statistics systems in university libraries[J]. Library and Information Service, 2016, 60(5): 51-56. [4] Yan Xiaodi, Shao Jing, Zhou Qi, et al. Design and implementation of electronic resource utilization statistics gateway system[J]. New Technology of Library and Information Service, 2018, 24(8): 97-100. [5] Wang Xiaoliang, Wang Wei. Construction of e-journal database statistics analysis system through firewall log mining[J]. New Technology of Library and Information Service, 2013, 29(z1): 122-126. [6] Guo Zhenying, Zhao Wenbing, Wei Yuhui. Analysis and design of electronic resource log statistics system[J]. New Technology of Library and Information Service, 2008, 24(9): 102-106. [7] Zhou Xin, Lu Kang. Research on reader behavior data mining based on library digital resource access system[J]. Journal of Modern Information, 2016, 36(1): 51-56. [8] Zhang Jilong, Yin Shenqin, Chen Tie. Research on library user information behavior data collection methods based on ERU: A case study of Fudan University Library[J]. Library Journal, 2014, 33(12): 10-16. [9] Zou Rong, Zhang Chengyu, Jiang Airong, et al. Design and application of electronic resource access management and control system[J]. Library and Information Service, 2010, 54(1): 121-124. [10] Shi Xiaohua, Qian Yin, Xie Rui. Design and application of university electronic resource access management and control system[J]. Application Research of Computers, 2011, 28(3): 1042-1045. [11] Wang Zhengjun, Dong Xiaomei, Yu Xiaoyi. Design and implementation of digital resource evaluation system based on bypass monitoring[J]. Library and Information Service, 2015, 59(9): 52-57. [12] Wu Qunhui, He Sheng, Feng Xinling, et al. Research on electronic resource usage statistics model for scientific research in university libraries[J]. New Century Library, 2017(11): 37-40. [13] Gourley D, Totty B. HTTP: The Definitive Guide[M]. Translated by Chen Juan, Zhao Zhenping. Beijing: Posts & Telecom Press, 2012. [14] Qu J. Introduction to three methods for decrypting HTTPS traffic[EB/OL]. [2018-01-22]. https://imququ.com/post/how-to-decrypt-https.html. [15] Usage of SqlDependency class[EB/OL]. [2018-01-22]. https://www.cnblogs.com/lanchong/p/7125400.html.

Design and Application of Electronic Resource Usage Statistical Analysis System Based on Fiddler Agent

Chen Guang

Fujian Institute of Research on the Structure of Matter, Chinese Academy of Sciences, Fuzhou 350002

Abstract: [Purpose/significance] This paper compares and analyzes the main methods for libraries in China to independently collect electronic resource usage statistical data, and solves the problem that existing technical methods cannot acquire electronic resource access information based on the Https protocol. [Method/process] The technology is improved on the basis of the existing collection mode based on bypass monitoring. Through policy routing and Fiddler agent, the statistics and analysis of electronic resource usage data based on Https protocol access is achieved. Based on this, the electronic resource usage statistical analysis system is designed and applied. [Result/conclusion] This paper solves the problem of collecting data on the usage of electronic resources based on the Https protocol, and provides a reference for other libraries to collect their own statistical data on the usage statistics of electronic resources.

Keywords: electronic resource, policy-based routing, usage statistics data, Fiddler proxy, bypass monitor

Note: Figure translations are in progress. See original paper for figures.