Application Research of Online Transaction History Extraction System Based on Lucene

Hui Zhang(✉)

Nanning University, Nanning, Guangxi, China
1362757861@qq.com

Abstract. With the rapid growth of the Internet economy, online transaction fraud occurs frequently and the victims are all over the country, which brings great difficulties to the public security organs in investigating and handling cases. This paper analyzes the history files generated by computer in the process of online transactions, and designs a history extraction system using Lucene technology. Through this system, the files can be retrieved quickly and accurately, the transaction records stored in the computer can be extracted, and the computer crime can be helped.

Keywords: Lucene · Online transaction · Computer forensics · History

1 Introduction

As online shopping has become a daily fashion of modern shopping consumption, there are many cases of online fraud. The perpetrator can easily cheat others' property by using computer network technology and means of online transactions, with only a small investment of human and material resources. Online fraud is often difficult to obtain evidence, and the way of manually searching for evidence in the computer has been unable to meet the needs of computer forensics. At present, the public security organs in China use the encase forensics system of Xiamen Meiya company for the forensics of computer crimes, but the encase system is too professional for ordinary non professionals to use. The public security organs' acquisition and analysis of online transaction history records are mainly manual retrieval on the client to collect data information about transaction records. This way is difficult to ensure the accuracy and integrity of document collection, thus affecting the efficiency of handling cases. Aiming at the historical records generated in the process of online transactions, this paper designs an online transaction record extraction system, which can quickly and accurately extract the local online transaction history records. It is hoped that through this system, the public security business department can quickly extract the criminal evidence of online transaction history, so as to effectively combat online transaction fraud.

J. Yen et al. (Eds.): ICBIS 2023, AHCS 14, pp. 209–216, 2023.
https://doi.org/10.2991/978-94-6463-198-2_24

2 Lucene Full Text Retrieval Technology

Lucene is a toolkit based on full-text search engine, and the source code of the toolkit is open to users [1]. The core technology is based on the architecture of full-text search engine. By scanning each word in the document, it points out the frequency and location of each word, and provides a relatively complete indexing function and query function. Lucene can be nested into various applications as the full-text index engine behind the application to establish a complete information retrieval library [2]. The process of Lucene full-text retrieval is shown in the Fig. 1.

Lucene has two main functions: indexing and retrieval.

(1) Create index

The documents to be indexed are processed by word analysis and word segmentation to get a complete word set. Create an index for each word in the word set, establish a full-text index library, and store the index library files in disk files [3].

(2) Search index

Retrieval index is the process of full-text retrieval through the establishment of index base. First, the query request is processed by syntax, and the query tree is obtained by syntax analysis; Then, after the index is read into memory, the index is searched by using the obtained query tree to get the retrieved result document; Thirdly, according to the acquaintance degree of the query request, the search results are sorted, and the query results are fed back and displayed. Lucene provides an API that developers can use to complete the information exchange of the index [4], but it is not a complete full-text retrieval application, but provides indexing and search functions for the application. If you want to realize the retrieval function of the system, you need to carry out secondary program development on the basis of Lucene to realize different scene requirements. Lucene search function is mainly composed of seven source code packages, which realize different functions respectively [5]. The specific function table is shown in the following Table 1.

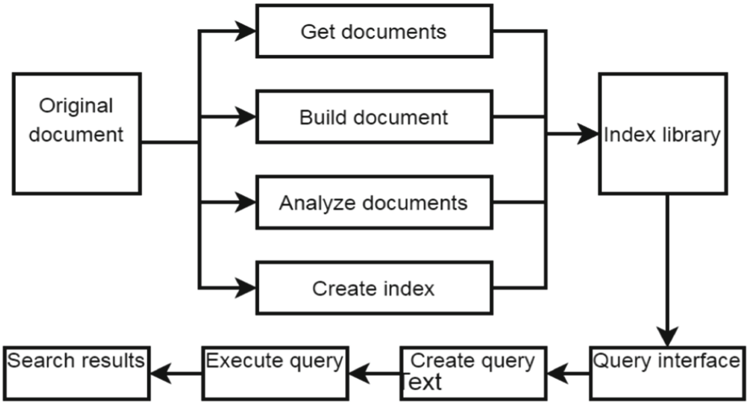

Fig. 1. Lucene's full-text retrieval process

Table 1. Corresponding menu of Lucene source code package

Source package	Corresponding function
Org.apache.lucene.Analysis	language analyzer, mainly used for word segmentation
Org.apache.Iucene.Document	structure management in document index storage
Org.apache.lucene.Index	index management, including index creation, deletion
Org.apache.lucene.QueryPaser	query analyzer, which realizes the operation of query keywords
Org.apache.Iucene.Search	search management, search results according to query conditions
Org. Apache. Iucene. Store	data storage management
Org. Apache. Lucene. Util	public class

3 Design of Online Transaction Record Extraction System

3.1 System Design Requirements

The online transaction record extraction system should meet the following requirements:

1. With local files as the search target, the system can automatically retrieve the online transaction history stored in the computer, and can also query the specific online transaction history through keywords.
2. It has high query accuracy and fast response speed.
3. Short search time and high efficiency.

3.2 System Function Design

First, establish a key Thesaurus of online transaction characteristics. When the system is running, use the keywords set in the key thesaurus to search and match the local web page file content. If you can't search the matching keywords, there is no local online transaction history file. If the user needs to further search and find a specific web page in the web page file of online transaction history that has been searched, the user can enter the corresponding keywords in the user interface, and the system will filter the search results that the user needs from the previous search results. The functional flow of the system is shown in the Fig. 2 below.

The system is composed of feature library and external interface. The feature library is the focus of the system. The feature library collects the web page features of online transaction records. The system can automatically find the matching content in local files through the keywords of the feature library. The builder can define the structure of the feature library. For example, it is necessary to load the characteristics of transaction records into the feature library in order to match the local web page file content and search the relevant web page content through the feature library. In the external interface, three functional modules of Lucene system are mainly used.

(1) Lucene source code package org.apache.lucene.analysis realizes the segmentation of relevant online transaction web documents, and serves as the language analyzer

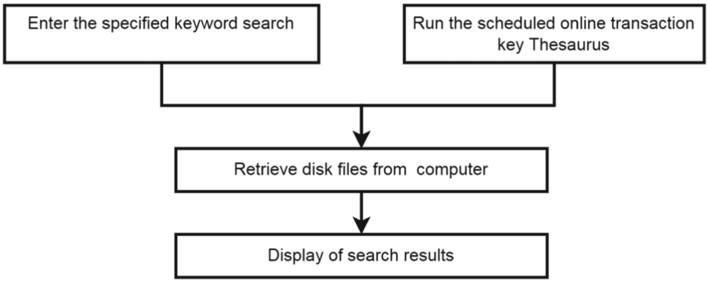
Enter the specified keyword search
Run the scheduled online transaction key Thesaurus
Retrieve disk files from computer
Display of search results

Fig. 2. System function flow.

of online transaction records. The work of word segmentation is realized by the extended class of analyzer. You can refer to the word segmentation analyzer class written by the implementation of this class to realize the functions of Chinese analyzer.

(2) Lucene source code package org.apache.lucene.search realizes the web page retrieval interface of online transaction records. Through the function interface, you can enter conditions to obtain the query result set, and you can also customize the query rules to realize the and, or, non and other composite queries of query conditions.

(3) Lucene source code package org.apache.lucene.store realizes the storage management of online transaction record related data, including some underlying I/O operations.

4 Web Page Features of Online Transaction Records

Feature extraction refers to how to extract valuable features from web pages to represent web pages. Now we extract and analyze the possible features of online transaction information records.

(1) Text information in the transaction web page. The first is the plain text content in the transaction web page. The second is the text in different fields in the transaction web page, including the web page title, web page subtitle, metadata description text, metadata keywords. For example, “Alipay”, “online payment is safe and fast!” Page title of.

(2) Mode of Trading Web page. Mode refers to a certain form of combination in the web page, for example, “¥ + number + 元” is a mode representing price. If there are some tables composed of prices in the web page, it can be preliminarily determined that this web page belongs to the “shopping” web page, and it can be determined that this is a transaction web page record.

(3) The link information page of the web page. Hyperlinks of web pages are an important difference between web pages and plain text, including web page category information with hyperlinks and web page text with hyperlinks. For example, there are some special link texts in the transaction process, such as“Click here to view the details of this transaction”, “Return to transaction management”, which is the link text information generated in online transactions.

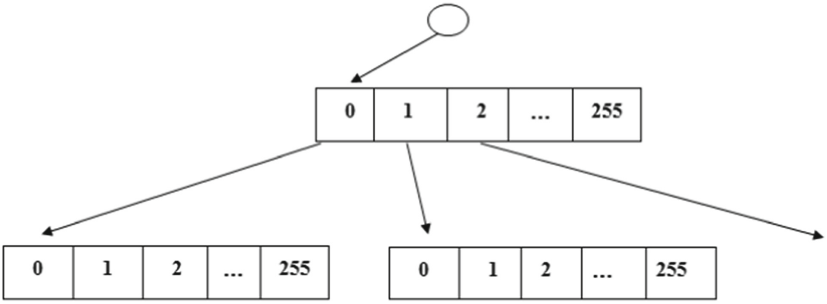

Fig. 3. Keyword tree.

(4) Use the URL information of the web page to represent the web page. Pages with the same URL prefix may belong to the same class of website pages. The longer the prefix of the same part of the URL, the more likely it is to belong to the same class of website pages.

5 Online Transaction Search Algorithm

5.1 Tree Based Multi Keyword Search Algorithm

Multi keyword text search refers to searching for the number and location of all keywords in an article with any length (byte string) according to the content of multiple keywords. The keywords are organized into a tree based structure. Starting from the root node, the strings on the path of the root node are connected together to form a keyword. All the keywords thus formed belong to the set of keywords to be searched. All keywords are stored in the tree data structure [6]. The retrieval process of keywords is a byte matching process from root node to leaf node, which improves the retrieval speed. The algorithm establishes each node of the tree in bytes. One byte is 8 bits, so the maximum number of nodes in each layer of the tree is 256 (28). The keyword tree is shown in the following Fig. 3.

5.2 Search Algorithm Design

The retrieval object is mainly for keywords, which are mainly composed of Chinese characters. The Chinese character coding generally adopts GB 2312 coding. The coding of a Chinese character is composed of 2 bytes. In the 8-bit information of each byte, the highest bit is l, while the high bit of ASCII code of other characters is 0 [7]. In view of the particularity of Chinese character coding, the system designs a tree based multi keyword fast search algorithm, and the algorithm flow chart is shown in the Fig. 4 below.

Each time the retrieval algorithm returns from the leaf node, it determines whether the returned text byte is a Chinese character byte. If it is an ASCII code that is not a Chinese character, the search will continue normally. If it is a Chinese character byte, the decision of whether to move forward is made according to whether the current byte is the first byte of the Chinese character, and it is guaranteed not to start from the second

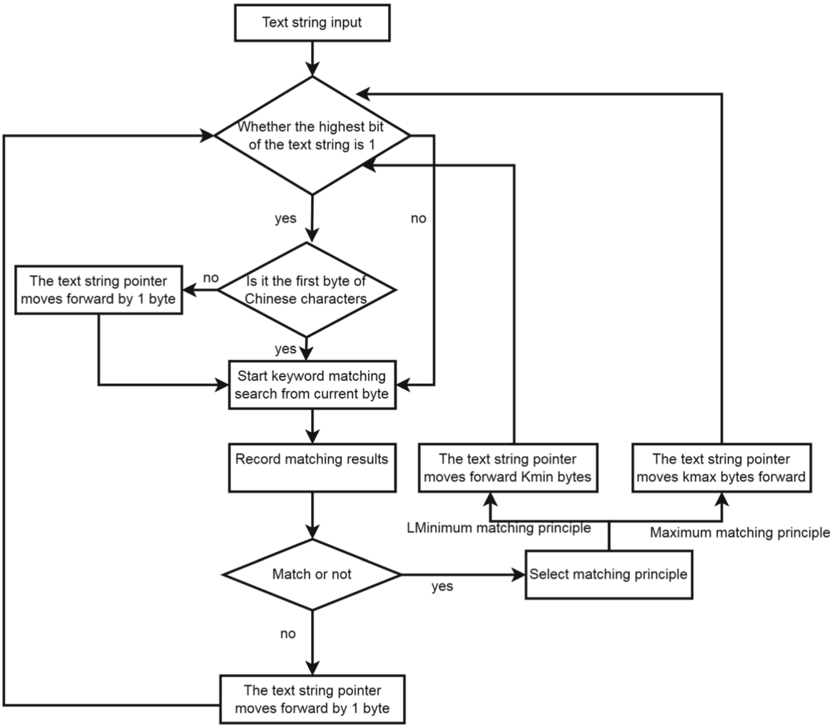
Text string input
Whether the highest bit of the text string is 1
yes
no
The text string pointer moves forward by 1 byte
no
Is it the first byte of Chinese characters
yes
Start keyword matching search from current byte
Record matching results
The text string pointer moves forward Kmin bytes
The text string pointer moves kmax bytes forward
LMinimum matching principle
Maximum matching principle
Match or not
yes
Select matching principle
no
The text string pointer moves forward by 1 byte

Fig. 4. Algorithm flow.

half byte of the Chinese character, because the search starting from the second half byte of the Chinese character will definitely not find a matching keyword, which belongs to useless search, so as to improve the search speed [8].

6 System Test

The system uses JDK development platform, Lucene toolkit interface and eclipse development. Enter the keywords “Alipay” and “online payment” in about 10000 document samples on the computer, and use Lucene system and MySQL system to search and compare them respectively. The data results are shown in the following Table 2.

Because the MySQL system does not preprocess the search words, there is a large difference between the hit number of MySQL query results and Lucene search results. The time-consuming of MySQL query is significantly higher than that of Lucene system, and the time-consuming of Lucene system is lower. The experimental data shows that the retrieval efficiency and accuracy of the system are significantly higher after the improvement and optimization of the retrieval algorithm combined with Lucene framework.

Table 2. Comparison of the number of hit documents and query time between Lucene system and MySQL retrieval (s)

Search keywords	Lucene		MySQL	
	Number of hits	Query time	Number of hits	Query time
Alipay	3301	1.76	3224	5.13
Online transaction	2679	1.56	2582	5.86

7 Summary

According to the actual needs of online transaction history forensics, this paper studies the application needs of full-text retrieval technology in computer crime forensics, and proposes an online transaction history extraction system. Based on Lucene, the system constructs a search engine for full-text query, realizes the automatic process of computer crime forensics, and improves the efficiency and accuracy of data extraction by improving the retrieval algorithm. The test shows that Lucene system search can accurately find the relevant information of online transaction records, with less time-consuming and reasonable search results.

References

1. Gao Jian Design of scientific and technological novelty retrieval method based on Lucene retrieval tool [J] Integrated circuit applications, 2022,39 (04): 114–115.
2. Zhang Chen, Chen Zhang Jian, Liu Jiangtao, Ren Fu, Zhang Hongwei Improvement and implementation of address matching method for Lucene adaptive word segmentation [J] Surveying and Mapping Science, 2021,46 (10): 185–193.
3. Jia Xiaoxia Lucene's dynamic information retrieval system design of network resource index [J] Microcomputer applications, 2021,37 (01): 55–58.
4. Luo Dongxia, Qing Lingbo, Wu Xiaohong Design and implementation of Chinese yes no question answering system based on Lucene [J] Information technology and network security, 2020, 39 (11): 74–78[1] Zhao Huijie, Wei Yongqi, Jiang Jincheng Research on the application of enterprise knowledge system based on configuration management and Lucene full-text retrieval [j]Shandong coal technology, 2020 (06): 206–208+212.
5. Xiong Anping, Li Chuangen, Cao Chunjiang Lucene index segment merging optimization strategy [J] Journal of Chongqing University of Posts and Telecommunications (NATURAL SCIENCE EDITION), 2020, 32 (01): 105–112.
6. Xu Aichun, Wei Yanhua, he Zhenmin. Design and implementation of e-government full-text retrieval system based on Lucene [J]. Modern information, 2008 (07): 223–225.
7. Li Guifeng Design and implementation of online transaction data mining and analysis system [D]. Shandong Normal University, 2014.
8. Wei Hua, Li Yangji. Application of tree based multi keyword search algorithm in network monitoring system [J]. Journal of Chengdu Institute of information engineering, 2005 (01): 81-83.