Journal of Networking and Telecommunications

ISSN: 2661-4065 Volume 2 Issue 1 (2020) http://ojs.piscomed.com/index.php/JNT

PM
Pisco Med Publishing

1

Editorial Board

Editor-in-Chief

Dr. Ankit Saxena
IET DAVV Indore India
India

Editorial Board Members

Dr. Ahmed Asal Kzar
The University of Kufa
Iraq

Dr. Hongchu Yu
Wuhan University
China

Dr. Alessandro Polo
University of Trento
Italy

Prof. Salvador Garcia-Ayllon
Technical University of Cartagena
Spain

Prof. Christos Bouras
University of Patras
Greece

Dr. Meisam Abdollahi
University of Tehran
Iran, Islamic Republic of

Dr. Deyu Lin
Nanchang University
China

Dr. Dimitris Kanellopoulos
University of Patras
Greece

Dr. Antonio Ruiz-Martínez
University of Murcia
Spain

Prof. Jae Eon Yu
Keimyung University
Republic of Korea

Prof. Dmitry G. Korzun
Petrozavodsk State University
Russian Federation

Prof. Aymen Flah
National Engineering School of Gabes
Tunisia

Dr. Zhenguo Gao
Huaqiao University
China

Dr. Kutubuddin Ansari
Kathmandu University
Nepal

Dr. Jiajun Shen
Zhejiang University
China

Dr. Sridhar Iyer
Jani College of Engineering
India

Dr. Abdulghani Ali Ahmed
Universiti Malaysia Pahang
Malaysia

Dr. Qian Yu
University of Regina
Canada

Volume 2 Issue 1 • 2020
ISSN: 2661-4065

Journal of Networking and Telecommunications

Editor-in-Chief

Dr. Ankit Saxena

IET DAVV Indore India

India

Pisco Med Publishing

Pisco-Med Publishing

2020 | Volume 2 | Issue 1

Journal of Networking and Telecommunications

http://ojs.piscomed.com/index.php/JNT

Contents

Original Research Article

Pisco Med Publishing

ORIGINAL ARTICLE

Analysis of Computer Network Security and Preventive Measures in the Data Age

Zhi Li*

Xi'an Aeronautical University, Lianhu District, Xi'an City, Shaanxi Province 710077, China. 252601586@qq.com

Abstract: With the development of science and technology, the network in human society continues to be improved and enriched, as well as the ability of database technology, which is of super cloud computing ability and extensive data sharing and brings great convenience to the whole country. But at the same time, the data that is used improperly in the network environment not only poses a certain threat to the network system, but also gradually erodes people's lives. For the computer network security problems in the era of big data, this paper first analyzes the location of the problems, and then puts forward the means to solve and prevent the security problems.

Keywords: Computer; Massive Data; Security Issues; Preventive Measures

1. Introduction

With the continuous development of science and technology, big data came into being. In the current network environment, big data has its obvious advantages, but there are also lawbreakers taking advantage of network loopholes, stealing the data information that should be correctly distributed, disturbing social life and seriously hindering the progress of the whole network security. Therefore, it is urgent to evaluate the security of computer network and give advice for the future development of network security, and the task of ensuring information security in the era of big data is arduous.

2. Internet security issues in the era of big data

2.1 Poor awareness of network security

At present, the most important problem is that people pay less attention to network security. Nowadays, the computer system has been equipped with a relatively perfect security environment. However, the low security awareness of network users themselves, such as setting simple passwords and disclosing security keys to others easily, makes the network security with an irreplaceable gap under strict monitoring[1].

2.2 Insufficient security of network software

Under the current conditions, with the help of network tools, people can better carry out daily life and work by using some network software. There are some problems in the design of software programs. When using certain software functions, they will be asked to fill in personal information with certain privacy. Users only believe in network software when considering the possibility of all network threats. Some program developers sacrifice part of security modules in order to enhance convenience, which also provides an opportunity for criminals to steal information. Once information is used, it will bring great damage to the personal life.

2.3 Computer virus and hacker flooding

The continuous development of computer technology is the most fundamental reason for the maximization of social convenience. With the continuous updating of database technology, there are also security problems in the network

Copyright © 2020 Zhi Li
doi:10.18282/jnt.v2i1.796
This is an open-access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

environment, which is the computer network virus and the network hacker who makes virus. According to different program security vulnerabilities and settings, viruses are all pervasive. With the characteristics of strong infectivity and adaptability, the network virus with changing forms gradually damages the network environment and steals personal information in daily life.

3. Internet security precautions in the age of big data

3.1 Strengthen the attention and supervision of enterprises on network security

For companies and related organizations that develop and deal with major issues through the Internet, it is far from enough to rely on the improvement of users' personal capabilities. A wide range and perspective of cognition is an effective way for companies and organizations to deal with network security issues. Enterprises should comply with the requirements of the era of big data, contact and help each other with various powerful network security protection platforms and institutions, timely update the internal database security precautions of the company, and timely optimize the current lack of network security regulations[2].

3.2 Apply monitoring procedures

For the protection of the network security environment, it needs the timely response of science and technology under the continuous and perfect supervision of the whole society. At present, the preferred approach for the maintenance of network security is a variety of security software systems and firewall technologies on the market. It is not enough to prevent the invasion and destruction of network virus for the updated computer network environment. In order to deal with the changes of virus after it attacks the computer, it is necessary to constantly give feedback for network prevention, that is, to feed back the successful network virus interception information to the big data, and to promote the use of the feedback information from the big data transmission by all kinds of anti-virus software in the society to predict the possible changes of the virus. With the help of the advantages of big data to deal with its shortcomings, this is how the development of science and technology progress can be achieved.

3.3 Improve network information transmission security

Although wireless transmission technology brings people a convenient way of computer network, there is a certain security threat. There is a virus program on the USB flash disk. Every time the user uses it on a different computer device, the virus will continue to spread endlessly. The only way to eliminate the virus will cause the loss of important data in the hard disk at the same time. However, as the wireless network technology corrodes the network environment in silence, only the link to the WiFi network will cause the personal information leakage. Through effective use of encryption technology on the network to transform the important personal information into the key, storing it in a large database, and at the same time improving the way to read the database encryption, people can protect personal data while using big data, and avoid it to be stolen in the process of data transmission[3].

4. Conclusion

The progress of computer network technology brings not only the convenience of life, but also some threats that are difficult to eliminate. In today's era of big data, it is necessary to improve personal awareness of network security, enhance the protection of network security, analyze and research while processing and prevention. Using the advantages of big data to gradually eliminate the threats and make up shortcomings brought by big data is the most effective solution and measures to protect the computer network security.

References

1. Hayes J, Courtney M. Hunter P. The protectors [Computer network security]. Engineering & Technology 2014; 9(5): 54-58. doi: 10.1049/et.2014.0520.
2. Wan X, Zhang L, Wei G, Hong H. Computer network security strategy research. Applied Mechanics and Materials 2014; 599(601): 1457-1460. doi: 10.4028/www.scientific.net/AMM.599-601.1457.
3. Chen H. Discussion on computer network security. Applied Mechanics and Materials 2013; 427(429): 2359-2363. doi: 10.4028/www.scientific.net/AMM.427-429.2359.

Pisco Med Publishing

ORIGINAL ARTICLE

Research and Application of Code Similarity Based on Submission

Yu Lang*

School of Information Engineering, Shanghai Maritime University, Shanghai 201306, China. E-mail: 850037409@qq.com

Abstract: With the continuous accumulation of resources, the similarity detection of code is becoming more difficult, and the difficulty of code reusing and rechecking is also increasing. In view of this problem, this paper proposes a code recommendation and check-research based on submission, which uses differential code cloning and word vector methods to find candidate code sets that are similar to incremental text, and uses feature extraction and clustering to select the most relevant codes from the candidate code sets to obtain repetitive codes. At the same time, it is recommended to programmers combined with relevance scores. Experimental results show that this method is feasible to some extent.

Keywords: Commit; Incremental Analysis; Code Similarity; Code Recommendation

1. Overview

In the field of software engineering, the efficiency of software development has always been a core concern of the software industry and academia. The concept of a „software crisis“ was proposed at a scientific conference organized by the North Atlantic Treaty Group in 1968[1], which still exists at present. During the development process, all developers hope to find existing software or code that meets their needs to help them save development time, so they spend a lot of time searching. In today's software development environment, with the widespread application of information libraries such as software version control libraries, a large amount of data information has been accumulated, providing a rich source of reusable open source resources for software reuse. However, these resources are huge, highly dispersed, diverse, and closely related to each other, which poses great challenges to the accurate positioning of reusable resources. Compared to the above active search code, code recommendation came into being.

In addition, with the development of network technology and software scale, software development works as a group, allowing different project developers to jointly participate in the collaborative development of the project. While improving efficiency, it also makes the software development process exist a lot of code redundancy. In order to achieve effective management and reuse of the code and help to rationally restructure the software, it is also necessary to add a process of repeatability detection in the version code.

2. Related work

Code similarity detection is widely used in various aspects such as check weight and code recommendation. The earliest code similarity detection was the attribute counting method, which was proposed by a foreign scholar Halstead[2]. However, the attribute counting method only considers the attribute characteristics contained in the code and ignores the code's organizational structure information, which makes the it often perform unsatisfactorily in the detection of "innovative reorganized code". In the latter development, detection methods based on structural metrics came out. In the algorithm of structural measurement[3], the focus is on the algorithm of code conversion into the identification string

Copyright © 2020 Yu Lang
doi:10.18282/jnt.v2i1.810
This is an open-access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

and the similarity measurement algorithm of the identification string. Although these studies are dedicated to helping developers better solve the problem of code reuse, most aspects of their work are biased towards API recommendations. This granular code completion recommendation method can provide less information. Therefore, the problem of low efficiency is common. In addition, some code completion recommendation methods are prone to problems such as the same or similar code being repeatedly recommended. In addition to research on code similarity, version-based hosting platforms have developed at home and abroad. Today, version management tools are combined with code management. It leads high time complexity of code similarity detection, indicating that it can no longer meet the urgent need to obtain the similarity of code change information. In response to these problems, in view of the characteristics of small submissions and frequent submissions by developers, this paper proposes research topics based on submitted code recommendations and rechecks, and uses an incremental analysis mode to re-analyze the version submissions. After analyzing the results, update them to the original results. In this way, the analysis scale can be effectively reduced, and the purpose of reducing the overall analysis time can be achieved.

3. Algorithm design

This paper designs a code comparison analysis method that uses the submitted code as the detection unit, which treats open source projects as a collection of submitted documents. First use Git and other command-line tools to extract the warehouse code, and then design the analyzer to filter out the difference code. In order to facilitate data storage and comparison analysis, the matching algorithm and word vector technology based on sliding window technology and mismatch index technology are used to determine the similarity relationship of code blocks, and perform duplication check or code recommendation. The design framework based on the submitted code recommendation and duplication research is shown in Figure 1.

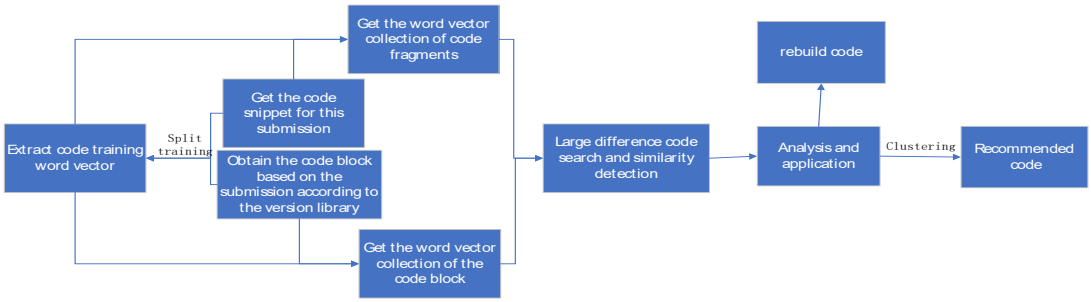


Figure 1. Design framework based on submitted code recommendation and duplication research.

3.1 Extracting code

Use the Git command line tool to obtain the open source warehouse code, use GitHub to save the open source project, and establish a connection with the remote warehouse. Design the analyzer, call Git log and other interfaces to get the commit log in the code warehouse, for each version snapshot, get the difference code between the commits, and traverse to get the code of the last commit version in the code warehouse. For the sake of conciseness, the difference code submitted above will be referred to as „incremental text“. The incremental text extraction algorithm steps are as follows.

1) Scan the incremental text log information and divide the log according to the submission.

2) Divide the divided commit log by documents, and get the incremental text of the source code documents.

3) Analyze the code in the incremental text, and add, delete, or modify the incremental text according to the three types of operations.

4) Add to the information list according to the added, deleted, modified code.

3.2 Preprocessing files

Some redundant information in the program may affect the vectorization of the program, such as comments, header files, spaces, and blank lines. Therefore, before vectorizing program features, it is necessary to remove redundant information in the program code that affects feature extraction. This process can not only make the extraction of program features more accurate, but also greatly reduce the source code file and speed up the operation efficiency of the

entire similarity detection algorithm[5].

3.3 Converting word vectors

In order to reduce the reliance on expert experience and avoid the problem of code similarity detection with similar semantics, word vector conversion is performed on the code. Treat all the text in the code base as a corpus, the sentences as words, and use the document as the context of the words. Use Word2Vec to calculate the word vector of each word in the text and convert it into each line of code Vector[6].

3.4 Similarity measure

Because the size of incomplete code is quite different from that of the complete code block, it is necessary to choose a method. The matching algorithm based on sliding window technology and mismatched indexing technology proposed in the research of classification of malicious code[6] can effectively solve this problem. It uses a sliding code window (i.e., continuous code snippets) instead of a single word vector as the basic unit of comparison. For each code window of length q, if each window allows e mismatches, extract all its substrings of length q-e. If there is at least one substring match, the two windows are considered a successful match. If the number of matching windows in the two code blocks meets the set similarity threshold, the two code segments are considered to be similar codes.

This paper improves the algorithm based on this algorithm[7], and the specific steps are as follows.

1) Because of coarse-grained filtering, the similarity threshold set in this paper is lower than the similarity threshold for general similar codes to find more alternative code block.

2) Since only similar code lookups are performed on submitted code blocks, this paper changed the search mode from a code-to-many mode to a one-to-many mode.

3) Since it is only necessary to find code blocks similar to the submitted code block, this paper does not store the key-value pairs of all sliding windows, but only the key-value pairs of the code block window to be completed, effectively reducing the storage space.

4) In order to finely divide the candidate code blocks found and improve efficiency, the method is to design and extract word vectors from the codes in the candidate code set, and to calculate word vector similarity.

5) Considering the distribution of the previously unknown code vectors in the data space, it is also unknown how many classes need to be divided. Therefore, this paper selects DBSCAN as the clustering method to finely divide the candidate code blocks, which is a density-based clustering The algorithm has a certain ability to resist noise. There is no need to set the number of generated categories. It only needs to set a similarity threshold. If the similarity between the vectors of two code blocks is less than the set threshold, they belong to one category.

3.5 Checking and recommending

Due to the need to search for duplicates and recommendations through similar code search methods, the author chooses the algorithm parameters more suitable for this problem through calculation experiments. This paper clusters the searched results (candidate codes) to make the same or similar code blocks get repetitive codes. Since there are a large number of duplicate codes in the code base, in order to prevent the same or similar candidate codes from being repeatedly recommended, the same cluster group is regarded as a type of recommendation. In addition, in order to ensure the usefulness of the recommended code, it is necessary to rank the candidate code blocks after clustering, so that the candidate code blocks with high relevance are recommended first.

4. Analysis of experimental results

In this paper, taking Python as an example ,the experiment randomly selects the https://github.com/donnemartin/system-design-primer open source project on Git Hub as the analysis object. The following experiment is designed, and the program comparison analysis is performed from the perspective of engineering similarity. The rationality of this method is verified through judging the code similarity.

As shown in Figure 2, using the word vector and the difference code similarity detection on the code snippets selected from the 25th code block, a similarity heat map as shown in Figure 3 is obtained. Among them, the similarity between the first code block , the 25th code block and the excerpt is 1, and the similarity of the 3rd, 5th, 6th, 19th, 21st,

and 23rd code blocks is 0.33. Experimental results show that this method is feasible to some extent.

```
def mapper(self, _, line):
    yield line, 1
def reducer(self, key, values):
    total = sum(values)
    if total == 1:
    yield key, total
def steps(self):
```

Figure 2. Excerpt from the code block.

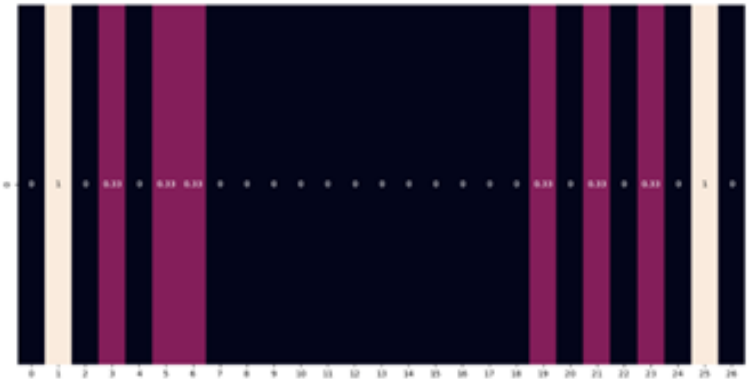


Figure 3. A similarity heat map of code blocks and excerpts.

5. Summary and prospects

With the continuous accumulation of resources, the difficulty of code reuse and duplication check is increasing. In response to this problem, this paper proposes a code recommendation and duplication check based on submission, which uses differential code search and word vectors to find alternative code sets that are similar to incremental text, and uses feature extraction and clustering to select the most relevant codes from the alternative code sets to obtain repetitive codes. At the same time, it is recommended to programmers combined with relevance scores. Experimental results show that the method is feasible to a certain extent. In the future, information such as abstract syntax trees will be used to further improve the accuracy and efficiency of this method.

References

1. Naur P, Randell B. Software engineering: Report of a conference sponsored by the NATO Science Committee, Garmisch, Germany, 7-11 Oct. 1968. Brussels: Scientific Affairs Division, NATO; 1969.
2. Halstead MH. Elements of software science. New York: Elsevier Science Inc; 1978.
3. Verco KL, Wise MJ. Software for detecting suspected plagiarism: Comparing structure and attribute-counting systems. Proceedings of the 1st Australasian conference on Computer science education; 1996 Jul; Sydney, Australia. 1996. p. 130-134.
4. Xu F, Hao L, Chen F, et al. A comparative analysis method for open source code reuse (in Chinese). Computer Engineering 2020; 46(1): 222-228+242.
5. Hu Z. Research and application of program code similarity detection method (in Chinese). Central South University; 2012. doi: 10.7666/d.y2197724.
6. Qiao Y, Jiang Q, Gu L, et al. Classification of malicious code based on assembly instruction word vector and convolutional neural network (in Chinese). Netinfo Security 2019; (4): 20-28. doi: CNKI:SUN:XXAQ.0.2019-04-004.
7. Yin K. Research on block completion recommendation algorithm based on differential code cloning search (in Chinese). University of Science and Technology of China; 2019.

Pisco Med Publishing

ORIGINAL ARTICLE

Design of Smart Card Chip Reader Based on STM32

Zebin Tan[1], Yong Zuo[1], Hongwei Cao[2], Jiahao Huang[3]

[1] School of Optical & Electronical Information, Beijing University of Posts & Telecommunications, Beijing 100876
[2] Baidu Online Network Technology (Beijing) Co.,Ltd., Beijing 100193
[3] School of Information & Electronics, Beijing Institute Of Technology, Beijing 100081

Abstract: This paper designs a smart card chip reader based on STM32, and gives a block diagram of the overall design of the system. It introduces the design of the peripheral hardware circuit of the main control chip STM32F103CBT6 and the RF processing chip THM3070, and the format of data communication between each part and flow chart. The communication between the host computer and the reader adopts the USB CCID protocol, which supports the USB full-speed communication rate of 12Mbps, improves the interrupt response, and overcomes the shortcomings of the traditional serial port-based readers with low communication speed and poor interrupt response. At the same time, it realizes data interaction with the PC/SC standard PC application and has better versatility.

Keywords: Circuits and Systems; Reader; STM32; USB CCID; THM3070; PC/SC

Introduction

Radio Frequency Identification (RFID) is a kind of automatic identification technology, which realizes non-contact two-way data communication through radio frequency 35 mode, identifies targets and obtains data[1]. At the same time, the identification of RFID does not need manual intervention, it can work normally in a relatively harsh environment. Radio frequency identification system is usually mainly composed of two parts: read/write module (i.e. card reader) and electronic transceiver (i.e. card). The card reader transmits a signal with a specific frequency rate through the antenna. When the card enters the radio frequency field of the antenna of the card reader, induced current will be generated, and the card will acquire energy and be activated.

The chip then sends the encoded information through the antenna built in the card, and the card reader antenna receives the carrier signal sent by the card number, then demodulate it, the demodulated and decoded information will be sent to the main controller, and finally the main controller will make corresponding processing according to different settings.

In the smart card issuing company, the traditional card writing method is to package the chip and antenna coil into a card with a plastic mold. After that, the card is initialized by means of antenna coil induction. Due to the complex electrical environment in the factory and the large radiation field of the card reader antenna during batch operation, the carrier signals emitted by different card readers interfere with each other, which may lead to the initial card.

In addition, some smart card chips are bad when they leave the factory. If these bad chips and antenna coils are packaged into cards with plastic molds before operation, the waste of coils and plastic molds will be caused and the production cost of enterprises will increase. Based on this, this paper designs a reader-writer that directly operates the smart card chip. At present, most of the card readers in the market are realized by single chip microcomputer and serial communication. The data transmission rate and interrupt response speed of this communication mode have great limitations. USB is an efficient, fast and economical serial communication interface, its ease of use and scalability has

Copyright © 2020 Zebin Tan *et al.*
doi:10.18282/jnt.v2i1.838
This is an open-access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

been widely supported and applied in the industry[2]. This design adopts USBCCID protocol as the communication mode between reader and PC, which has higher transmission efficiency and response speed than traditional serial communication, and is compatible with PC application program conforming to PC/SC standard.

1. Overall design of system

The card reader designed in this paper mainly includes: (1) the control circuit with STM32 controller as the core, and the corresponding clock 55 circuit and power supply circuit; (2) SPI communication magnetic coupling isolation circuit with adum3151 as the core; (3) The smart card chip data processing circuit with THM3070 as the core includes corresponding filtering and envelope detection circuits. As shown in Figure 1. The STM32 controller controls the whole system. It can receive commands from the upper computer and send them to after analysis.

THM3070, at the same time it will also THM3070 feedback information back to the upper computer. The main work flow of the system is as follows: the card reader is connected to the upper computer through the USB interface, and the USB module in the STM32 at this time. The block starts to enumerate the card reader as a CCID device, and then the controller will wait for the upper computer program to send instructions. When it receives the APDU command from the host computer, it will analyze it according to the format of the CCID protocol. The analyzed commands are transmitted to the radio frequency processing chip THM3070 through the SPI protocol, and then the THM3070 converts these commands into radio frequency signals and sends them to the smart card chip. The response signal returned by the smart card chip is sent to THM3070 after envelope detection. Then it is transmitted back to STM32 through SPI interface. STM32 packages the information into CCID protocol format and uploads it to the upper computer.

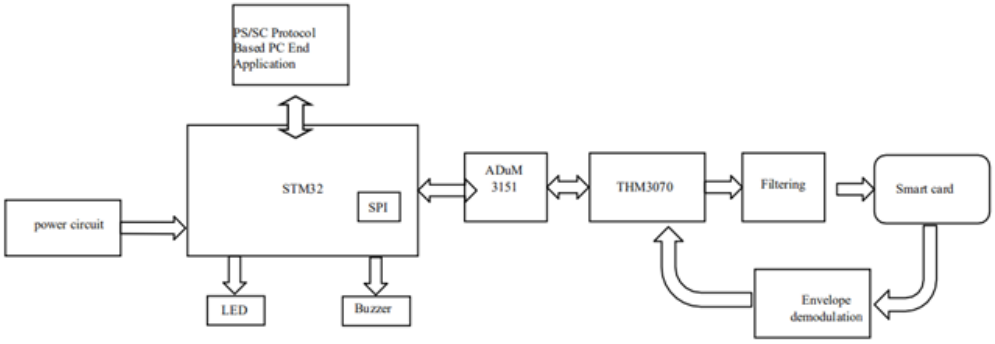


Figure 1. System overall design.

2. Hardware circuit design

2.1 Main controller and its peripheral circuits

2.1.1 Master circuit

In this design, STM32F103CBT6 is used as the main control chip to complete the scheduling of all interfaces and the processing of events. 75 STM32 is a 32 bit single chip computer designed by stmicroelectronics Company. It is based on the ARM Cortex-M3 core, which reduces the power consumption of the system and has the characteristics of high performance and low cost. The highest operating frequency of STM32F103CBT6 is 72MHz with fast interrupt response capability. The SPI interface has a full duplex and half duplex communication rate of 18 Mbit/s in slave or master mode. The 3 bit prescaler can generate 8 main mode frequencies and can be configured as 8 bits or 16 bits per frame. CRC generation/verification of hardware supports basic SD card and MMC mode, and all SPI interfaces can use DMA operation. STM32F103CBT6 is embedded with a USB device, which follows the USB full-speed standard and can realize full-speed (12Mbps) device functions; It has software configurable endpoint and standby /recovery function; The dedicated 48MHz clock is directly generated by the internal master PLL.

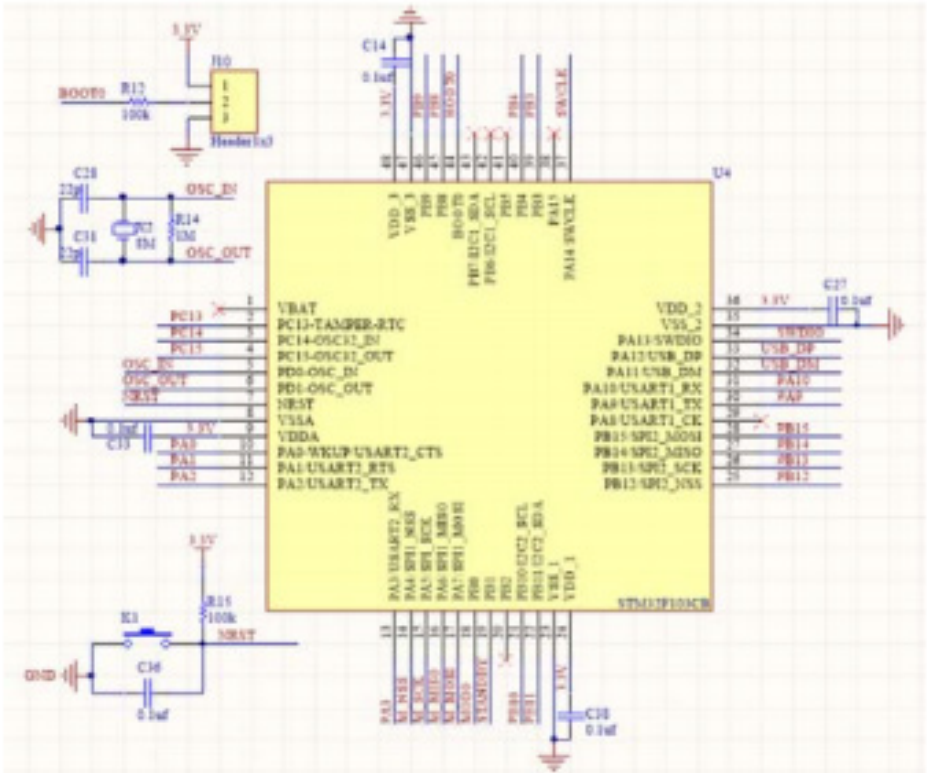


Figure 2. Main controller circuit.

2.1.2 SWD interface circuit

SWD (Serial Wire Debug) means serial debugging. Through this interface, program downloading and debugging of the chip can be realized. Compared with 20 pins of JTAG, SWD only needs 4 pins and occupies less GPIO ports.

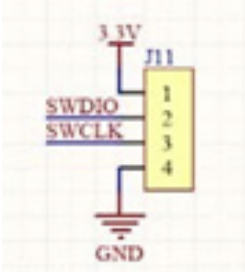


Figure 3. SWD interface circuit.

2.1.3 Buzzer and LED circuit

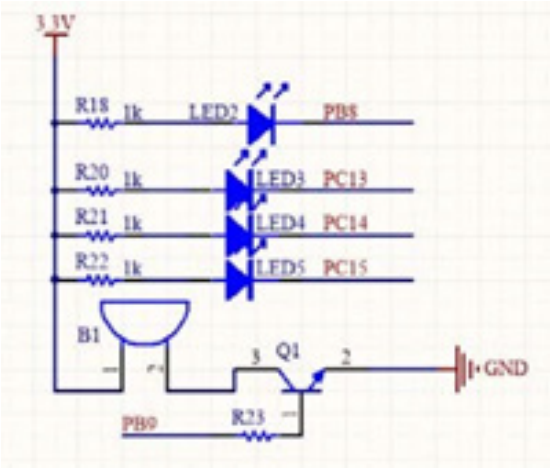


Figure 4. Buzzer and LED circuit.

2.2 THM3070 and its peripheral circuits

2.2.1 THM3070 circuit

THM3070 is a smart card read-write chip working at 13.56 MHz, with built-in power amplifier driver and adjustable transmission power. It conforms to ISO/IEC14443 TypeA/B and ISO/IEC 15693 standards, supports data transmission rates of 106 kbit/s, 21 2kbit/s, 424 kbit/s and 848 kbit/s, and has a maximum length of 256 bytes. The communication interface between the external controller and THM3070 has SPI mode or IDR mode, in which IDR interface can only be used for the second generation ID card security module. This system mainly uses SPI interface. The external controller communicates with THM3070 through SPI protocol to realize read-write memory sender, read-write data, send-receive control, baud rate control and protocol selection.

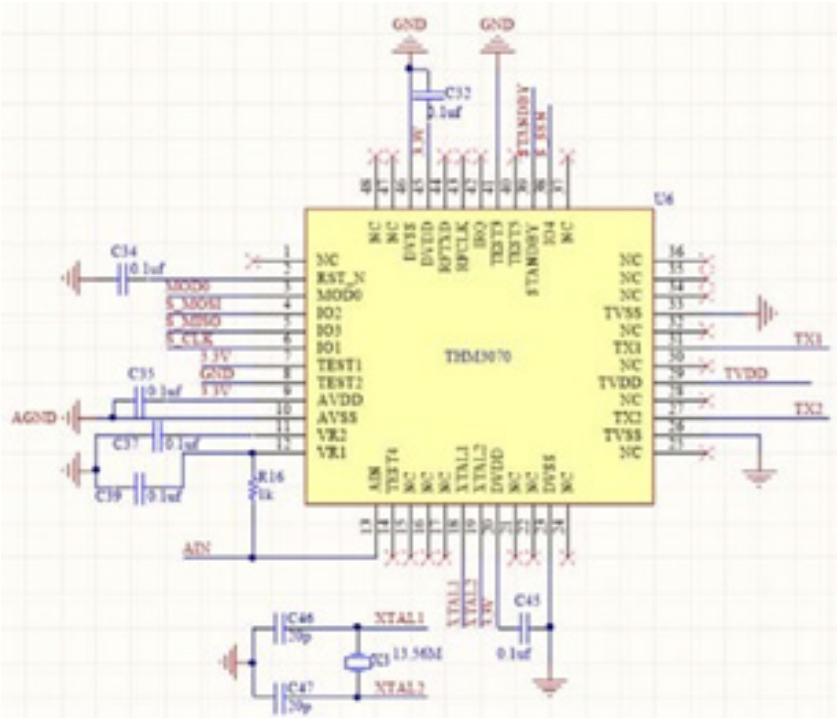


Figure 5. THM3070 circuit.

2.2.2 Filter and signal transmission circuit

As shown in Figure 6, TX1 and TX2 are respectively connected to the 31 and 27 pins of THM3070, which are the outputs of the chip power amplifier pins. The output signals are filtered by L1, L2, C15, C16, C21 and C22 and then contact with the pins of the smart card chip through the J14 interface.

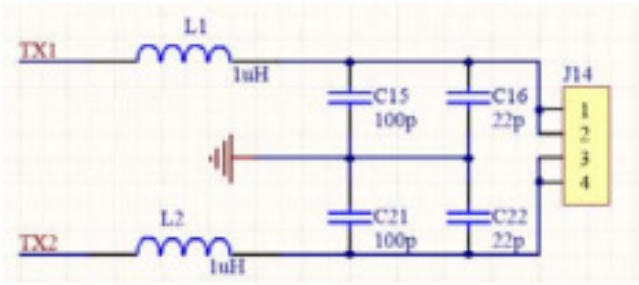


Figure 6. Filter and signal transmission circuit

3. Related program design of reader

In this design, the communication between the main controller and THM3070 is realized by SPI protocol. SPI (Serial Peripheral Interface-Serial Peripheral Interface) allows MCU to communicate and exchange data with various

peripheral devices in serial mode. SPI bus usually consists of 4 lines, namely serial clock line (SCK), master output and slave input line (MOSI), master input and slave output line (MISO) and slave select line (SSN)[3]. The THM3070 SPI interface supports clock normal low, rising edge valid timing, and SSN to remain low within one frame of data.

When writing register data, STM32 first issues the address to be written and then follows the data to be written. During the operation period. SSN must remain low, and THM3070 samples data on the rising edge of each clock. When reading register data, STM32 issues the address to be read by MOSI, MISO outputs the data in the corresponding address, THM3070 outputs the data on the falling edge of the clock, STM32 samples the data on the rising edge of each clock.

When writing buffer data, the write address is fixed at 0x80, and the data immediately following it cannot exceed 256 bytes. During operation, SSN needs to be kept low at all times. When reading buffer data, the read data address is fixed to 0x00 and SSN is kept low during operation.

4. Conclusion

This paper introduces in detail the hardware and software design of the reader-writer based on STM32 smart card chip. The main controller adopts STM32F103CBT6. The chip supports full speed USB communication rate of 12Mbps, which improves the data transmission rate of the host computer and reader-writer. The card reading chip adopts THM3070 of Uni-Light Co., Ltd. which supports 280 ISO/IEC144443 TypeA and TypeB standards and ISO/IEC15693 standards. In view of the disadvantages of low speed and poor interrupt response when traditional reader-writer uses serial port to communicate with host computer, this design adopts USB CCID protocol to re-package the communication data packets between host computer and reader-writer, and realizes data interaction between host computer application program and reader-writer conforming to PC/SC specification.

References

1. Zhou Xiaoguang, Wang Xiaohua. Principle and application of radio frequency identification (RFID) technology. Beijing: People's Posts and Telecommunications Press; 2006.
2. Chen Qingsong, Wang Jian. Universal serial bus data transmission. Computer Engineering and Design 2006; 27(11): 2077-2079.
3. Yang Meigang, Li Xiaowen. SPI interface and its application in data exchange. Communication Technology 2007; 40(11).

Pisco Med Publishing

ORIGINAL ARTICLE

Research on Telecom Fraud Detection Model Based on Cellular Network Data

Kaiyuan Guo, Wenbo Wang

School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876

Abstract: With the rapid development of wireless communication technology, the use of mobile phones and other means of communication for telecommunications fraud has become a major problem that endangers user security. Aiming at this problem, this paper constructs a telecom fraud user detection model by in-depth analysis and mining of cellular network data. The model includes data processing, CNNcombine algorithm and model evaluation. First, in the data processing part, the data set is subjected to feature screening, coding, sampling, and the like. Secondly, the CNNcombine algorithm is a combination of a one-dimensional convolutional neural network and multiple traditional classification algorithms. The convolutional neural network is applied to solve classification problems other than text image signals. Finally, in the model evaluation part, it is proved that the CNNcombine algorithm has higher accuracy than the common machine learning classification algorithm such as XGBoost to detect telecom fraud users.

Keywords: Machine Learning; Cellular Network Data; Deep Learning; Classification Algorithm

Introduction

Telecommunications fraud refers to the criminal act of criminals sending false information, setting up fraud schemes, and conducting long-distance non-contact fraud on the victims by means of telephone voice, short messages and other means to induce the victims to pay or transfer money to the criminals, which affects the social security and stability for a long time. Telecom operators have a large amount of data, the data scale is far larger than that of other industries data, this paper focuses on how to analyze and model cellular network data, and use this information to realize the detection of telecom fraud.

1. Detection model framework and data preprocessing

1.1 Fraud user detection model framework

The research point of this paper is to build a detection model of telecom fraud based on the measured data of cellular network provided by operators, and analyze electricity letter cheats the user's characteristic, finds the telecommunication cheats the user, and promotes the restriction measure regarding the telecommunication cheats the user.

1.2 Data preprocessing

1.2.1 Data set introduction

The cellular network data used in this paper are collected from a telecom operator in a city in Guangdong Province. Data has been desensitized, ODS.

Copyright © 2020 Kaiyuan Guo *et al.*
doi:10.18282/jnt.v2i1.835
This is an open-access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Package information: package information refers to package information handled by the user, including package price, package type and package name, totaling 3 characteristics.

Terminal information: information based on the user's terminal model, including terminal model, terminal category, terminal brand, etc. with a total of 9 features.

Internet surfing behavior: the usage of traffic collected from users, including daily average traffic, and the proportion of active days of traffic totaling 6 features.

Call behavior: including daily average calling times, daily average calling duration, calling duration 5 seconds, 5 seconds -15 seconds accounting for all the main time periods

The proportion of calls and the proportion of calling calls in each time period are based on the characteristics of the user's call behavior, totaling 33.

Short message behavior: The number of short message objects and the number of up-and-down messages are recorded with 5 characteristics.

Base station data: the average number of calling base stations per month, the average number of cities in which the called are located per month, and the proportion of calls made in areas with high fraud incidence, etc.

1.2.2 Feature selection

The purpose of feature selection is to screen the model entry indexes and to reduce the training complexity. The main basis for feature selection based on the features of samples is as follows.

(1) Features are less important

Percentage of missing values: attribute missing values exceeding 50% are rejected as invalid features. Category Proportion in Typed Features: Calculate the proportion of category values in the total number of typed variables; if the proportion exceeds 50%, the feature is regarded as unimportant and the feature is eliminated. Proportion of categories in classification features: calculate the proportion of each category in classification variables to the total number; If the percentage of categories is greater than 90%, the feature will be regarded as unimportant and the feature will be eliminated. (2) Whether the characteristic variable is independent of the label variable: For classification problems, features independent of labels are irrelevant features, and the purpose of feature selection is to remove irrelevant features.

The specific method of 90 is related to the type of index: for classified indexes, chi square test is a commonly used method in statistics to evaluate whether two events are independent. through chi-square test, the probability of variable independence p can be obtained. the smaller the value of p, the more important the characteristic variable is. in this paper, the characteristic with p value less than 0.05 is regarded as an important characteristic. For numerical indicators, the F test is a kind of hypothesis test method based on the F distribution. It analyzes whether the average value of the indicators has significant difference under different values of target variables, and the variables with significant difference have strong correlation. The independent probability P of variables is obtained through the F test. If the value of P is less than 0.05, the feature is related to the label and is an important feature.

1.2.3 Coding, sample sampling and division

(1) One-Hot Encoding classifier has discrete features that are not easy to solve, and cannot deal with problems such as character strings. It needs encoding. This article adopts the independent method. Thermal coding is used to solve these problems.

Single-hot coding is also called one-bit valid coding. It mainly uses N bit status registers to code N states, each state is separated by its own register bits, and only one bit is valid at any time. The values of discrete features extend to European space and the eigenvalues of discrete features correspond to a point[1] in European space. For example, the auto-likelihood code is 000,001,010,011,100,101 and the single heat code is 000001,000010,000100,001000,010000,100000. Each features become mutually exclusive, making the distance calculation between features more reasonable.

(2) Sample sampling

The number of positive samples in the original sample is about 699,000, the number of negative samples is about 18,000, and the ratio of positive and negative samples is 1:37.8. In order to improve the modeling efficiency, the original samples should be sampled and the amount of modeling samples should be controlled. At the same time, in order to

improve the accuracy of the model, the proportion of positive and negative samples for modeling is generally controlled to be around 1:5. The model randomly samples positive and negative samples respectively. The model used in this paper randomly samples positive and negative samples respectively, and selects 67,000 samples into the model, in which the ratio of positive and negative samples is about 1:5.

(3) Data Set Partitioning In order to ensure the generalization ability of the model, the data set needs to be divided into training set and test set. In this paper, the data set is randomly partitioned points, with 70% as the training set and 30% as the test set.

2. CNN combine algorithm

The data set used in this paper contains a certain number of positive samples. Therefore, the whole detection model is designed based on the theoretical tools of classification algorithm in machine learning. This paper improves the classification model in machine learning based on convolution neural network.

2.1 Classification algorithm based on one-dimensional convolution neural network

2.1.1 One-dimensional convolution neural network

Convolution neural network is a kind of depth feed forward neural network, which is generally composed of convolution layer and pooling layer alternating with each other[2] Have Characteristics of local connection, weight sharing and sub sampling. These characteristics make the convolution neural network invariant to translation, scaling and distortion to a certain extent, and the convolution neural network needs fewer parameters, so it has higher training efficiency. The convolution neural network updates its weight by using back propagation algorithm.

One-dimensional convolution neural network (1D-CNN, 1D Convolutional Neural Network) has the same principle and characteristics as two-dimensional or three-dimensional convolution neural network. The key difference is that the input of 1D-CNN is a one-dimensional vector, and the convolution kernel and feature map in the neural network are also one-dimensional. Convolution neural network is usually used in the fields of image processing, natural language processing and voice frequency/speech processing. In this paper, the cellular network data set is innovatively used as the input of convolution neural network, the characteristics of each user are regarded as a one-dimensional vector, and the two-classification function can be realized through the output layer of 1D-CNN.

2.1.2 Neural network structure design

The CNN convolutional neural network designed in this paper consists of input layer, 5 hidden layers, and output layer, as shown in Figure 1.

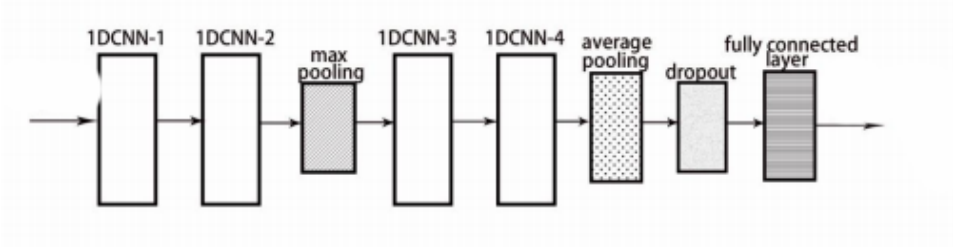


Figure 1. CNN convolutional neural network structure.

The first layer is the input layer: according to the data preprocessed previously, each data has a total of 63 features, so it is necessary to a vector with a length of 63 is transferred to the neural network, so here each user data needs to be reshaped to form A 21 x 3 matrix is used as the input of the neural network. The next 2-7 layer is the hidden layer and is the most important component of convolutional neural network.

The first 1D-CNN layer: defines a convolution kernel with a height of 2. The filter will allow the neural network to learn a single feature in the first layer, defining a total of 50 filters, which allows 50 different features to be trained on the first layer of the network. The output is 20×50 neuron matrix. Each column of the output matrix maintains the weight of a single filter. Use the defined kernel size and test. Considering the length of the input matrix, each filter will contain 20 weights.

Second 1D-CNN layer: single-layer convolution can only obtain shallow abstract information. In order to increase the depth of the neural network, a 1D-CNN layer is added. Results from the first CNN will be sent to the second CNN layer. Define 50 different filters again and train at this level. The second layer follows the same logic as the first layer, and the final output matrix size is 19 x 50.

Max Pooling Layer: pooling layer is usually used after convolution layer. The purpose of pooling layer is to reduce the complexity of output and prevent data over-fitting. The principle of maximizing the pooled layer is to take the largest eigenvalue as the value of the region in the neighborhood and set the pooled window size to 3. This means that the output matrix of this layer is only one-third of the input matrix.

Third and fourth 1D-CNN layer: add two convolution layers again to learn more global features, convolution kernel is large small is 2 and the number of filters is 80. The output matrices of these two layers are 5x80 matrix and 4x80 matrix. After a total of four convolution layers, we obtained neurons containing global features and realized feature extraction.

Global Average Pooling Layer: add another pooling layer to further avoid over-fitting. The global average pooling layer performs more extreme types of dimensionality reduction, and each feature detector has only one weight remaining in the neural network on the layer. The size of the output matrix is 1280 neurons, which is a one-dimensional vector and can be regarded as a feature vector after neural network feature extraction. It also provides the possibility to combine with other classification algorithms.

Dropout Layer: add a drop layer to solve the problems of overfitting and gradient disappearance. The discarding layer discards some neurons in the network on-board and sets the weight of 50% neurons to zero. The discarding layer averages and reduces the complex co-adaptive relationship between the divine elements, thus achieving the effect of over-fitting. The output of this layer is still the 1 x 80 divine element matrix.

Finally, the output layer is connected to the convolutional neural network in the form of a Full Connected. Since there are only two types of prediction targets: telecom fraud users and non-telecom fraud users, the vector with a height of 80 is reduced to a height of 2 at the last layer, which is completed by matrix multiplication at that layer. Using Softmax as the activation function, it can force the sum of all 2 outputs of the neural network to be one. The output value will represent the probability of each of the 2 classes.

2.2 Principle of CNN combine algorithm

In order to obtain better classification accuracy than 1D-CNN and traditional classification algorithms, we propose a classification model CNN Combination, which CNNcombine the 1D-CNN neural network designed in 2.1 with traditional classification algorithms. The advantage of 1D-CNN is that the convolution layer can extract high-level features from the input feature sequence and use the extracted features to improve the accuracy of the traditional algorithm model.

In addition, the prediction result of a single traditional algorithm model under the best parameters is the best performance of the model. Different parameters adopted by different models, in many cases, the prediction result of a certain model is not the best[3] but the prediction effect of the model on some samples in the sample set may be better than the original best prediction model. That is, the prediction results of each model are different, and it is possible to obtain better prediction accuracy by means of complementation. Therefore, a better model can be obtained by means of model fusion.

In this paper, the traditional algorithm for model fusion is to fuse multiple classifiers through meta-classifiers. After the secondary classifiers are trained, the prediction results are given based on the training data, and the meta-model is trained again based on the output of the prediction results of the secondary classifiers. Secondary classifiers can adopt many different classification algorithms, which can improve the generalization ability of the model and also reflect the advantages of each secondary classifier in different data. Therefore, the model is heterogeneous after fusion. The data set is divided into 4 parts, of which three are training sets and one is test sets. Adaboost[4], random forest[5], GBDT[6] and XG Boost[7] are integrated learning algorithms. Individual algorithms have achieved better prediction results on cellular network data sets. Therefore, the paper takes these four algorithms as the basis of integration. Firstly, AdaBoost, random forest and GBDT are selected as secondary classifiers, and the training set is subjected to K folding and cross training,

and a new data set is constructed according to the labels of the output results as new features of the data set. Then select XGBoost as the meta-classifier, train the meta-classifier according to the new data set, and finally output the prediction results of the comprehensive model.

The implementation method of combining 1D-CNN is to replace the last full connection layer in 1D-CNN with the classifier of comprehensive algorithm. The cellular network dataset is subjected to feature extraction by the neural network, and the length of the feature vector input into the synthesis algorithm is 80.

3. Model evaluation

In this section, data sets from telecom operators are used to verify the effectiveness of the model in detecting telecom fraud numbers. Experiments are designed to compare the performance differences between the proposed algorithm and the commonly used traditional classification algorithms.

3.1 Experimental design

The data set is randomly divided, with 70% as the training set and 30% as the test set. Use this dataset in Python. The training of classification algorithm is carried out under the environment. the selected algorithm includes the CNN combination classification model and design proposed in this paper.

The probability of being a telecom fraud user, the greater the value of the prediction result, the greater the probability that the sample belongs to a positive sample, i.e. a telecom fraud user, whereas the smaller the value, the greater the probability of being a normal user. Different thresholds can be set according to different requirements, and whether a sample belongs to a positive sample can be judged by comparing the results of the classification algorithm with the thresholds. The P-R curve (precision-recall)[8] and the average precision AP (average precision)[8] were used as evaluation criteria for the experiment.

In the P-R curve, P indicates the accuracy rate and R indicates the recall rate. With p as the abscissa and r as the ordinate, for the same model, a curve can be obtained according to the method of dividing points by taking samples one by one as the threshold value, which is the P-R curve. Assuming that the P-R curve of one model can completely surround another learner, the former has better performance.

3.2 Performance evaluation results

Compares the classification results of four different classification models, and the drawn P-R curve is shown in Figure 3 and Figure 4.

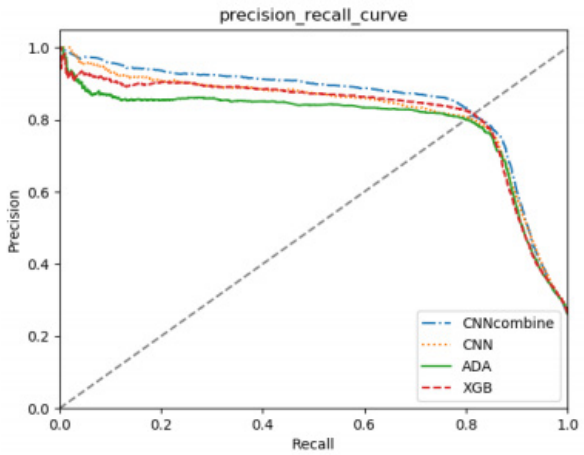


Figure 3. CNN combine algorithm structure.

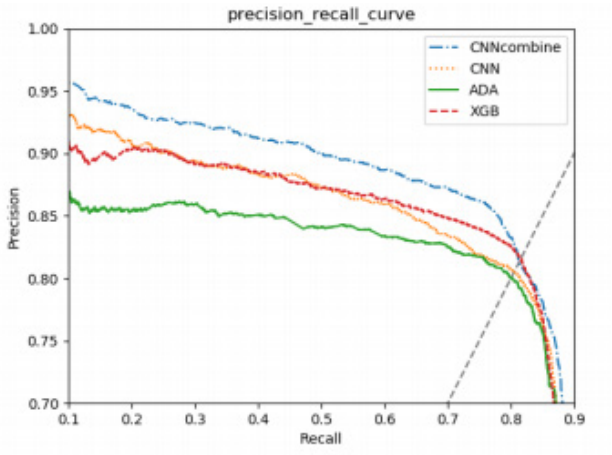


Figure 4. P-R curve comparison (upper right corner).

Comprehensive evaluation of the above model results can prove that the improved algorithm CNN Combination has better classification performance than other classification models in detecting telecom fraud users. On the other hand, when the 1D-CNN neural network designed in this paper is directly used to deal with the problem of detecting telecom fraud users in reality, the effect has no obvious advantage over traditional classification algorithms such as XGBoost. The actual evaluation proves that the telecom fraud detection model of CNN combine algorithm can more accurately detect fraudulent users from multiple users.

The output result of the telecom fraud detection model is the probability that the user is a fraud user. In practical application, it is necessary to set a classification threshold to determine the prediction result, i.e. those with a prediction probability greater than the threshold are classified as fraud users, and those with a prediction probability less than the threshold are classified as normal users. Generally, the threshold is determined by selecting the threshold corresponding to the same precision rate and recall rate, which can maintain a high precision rate and recall rate at the same time. When the precision rate and recall rate are the same, the classification threshold of the test set detection model is 0.41777, the precision rate and recall rate of the model are 0.8155, and 4275 fraud users can be correctly detected among 5242 fraud users in the test data set, and the classification accuracy rate for all 20133 users in the test set is 90.394%.

4. Conclusion

The content of this paper is to design and implement a detection model for telecom fraud users, and to propose a CNN combine algorithm that combines one-dimensional convolution neural network with traditional classification algorithm. Based on the experiment of the measured data of cellular network, this paper verifies that the CNN combination algorithm has better prediction results than the traditional algorithms XGBoost and Adaboost. The accuracy and recall rate are obviously improved. The average accuracy is 3% higher than XGBoost and 6% higher than Adaboost. This proves the effectiveness and feasibility of the model proposed in this paper in solving practical problems.

References

1. Xlinx. HDL Synthesis for FPGAs design guide[M]. XACT 1995: 3-13
2. Takahashi N, Nishi T, Hara H. Analysis of signal propagation in 1-D CNNs with the antisymmetric template[A].12th International Workshop on Cellular Nanoscale Networks and their Applications (CNNA 2010)[C]. Berkeley:IEEE.2010.
3. Xiao JZ, Lei B, Wang CQ. Reclamation on building waste produced from Wenchuan Earthquake[A]. Shanghai: Tongji University Press; 2008. 64-65.
4. Sill J, Takacs G, Mackey L, et al. Feature-weighted linear stacking[J]. Computer Science 2009.
5. Yoav Freaud, Robert E. Schapire. Decision-theoretic generalization of on-line learning and an application to boosting[J]. Journal of computer and system sciences 55: 119-139
6. L B. Random Forest [J]. Machine Learning 2001; 45: 5-32.
7. Friedman J H.Greedy function approximation: a gradient boosting machine[J]. Annals of Statistics 2001: 1189-1232.
8. Chen T, Guestrin C. XGBoost: A scalable tree boosting system[J]. In Proceeding KDD ,16 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco 2016; 785-794.
9. Zhou Zhihua. Machine Learning: Machine learning[M]. Beijing: Tsinghua University Publishing House; 2016. p.26-48.

Pisco Med Publishing

ORIGINAL ARTICLE

Research on Topoloty Reconstruction Mechanism Based on Traffic Identification

Qishuang Zhu, Hongxiang Guo, Cen Wang, Yong Zhu

Beijing University of Posts and Telecommunication School, Optoelectronic Information College, Beijing 100089

Abstract: Due to the growing variety of data center services, the bursty and variability of data traffic is increasing. In order to make the network better meet the needs of upper-layer services, it is necessary to design a more flexible optical internet topology reconstruction mechanisms to adapt the changing traffic demands. In the past research on optical internet, all topology reconstruction mechanisms are designed based on global data traffic. Although these mechanisms can fully utilize the flexibility of the data center optical interconnection network topology and adjust topology in real time according to the traffic demands, but when the traffic is presented at the regional level, this mechanism does not give optimal results. This paper proposes a topology reconstruction mechanism for data center optical interconnection network based on traffic identification for the previously proposed data center optical switching architecture—OpenScale. The simulation results show that it utilizes the flexibility of the network to save bandwidth resources and increase the wavelength connection bandwidth utilization with a little sacrifice of throughput.

Keywords: Optical Communication; Traffic Pattern; Area Reconstruction

Introduction

In recent years, with the rapid development of cloud computing technology, the traffic in the data center (DC: Data Center) has exploded, and the introduction of optical switching has become inevitable. Various new data center optical interconnection network architectures, such as Helios[1] OSA[2] and C-through[3], etc., usually adopt optical circuit switching (OCS: Optical Circuit Switching) to transport large amounts of traffic. Therefore, in order to enable the network to better match the various communication modes of the data center, a fast and high effect optical path reconstruction mechanism is essential. In our previous work, we proposed a data center optical interconnection network architecture based on small world, named Open Scale[4], in which a plurality of hexagonal rings form a regular lattice structure. In each hexagonal ring, nodes provide logical full connection in the form of optical burst switching. Each node also has the capability of wavelength exchange and can establish a direct reconfigurable wavelength path with remote nodes. The reconfiguration logic topology referred to in this paper is OCS wavelength The survey of data center traffic shows that the changing service requirements of the data center lead to different traffic characteristics[5,6]. Traffic patterns in data centers are correlated with applications, and they can be divided into two categories: regional traffic patterns and full traffic patterns.

A typical way to generate regional traffic patterns is to run multiple concurrent parallel computing jobs on the data center.Using the flexibility of the OpenScale network, the b-matching algorithm[9] can be adopted. this algorithm aims to preferentially select the node combination with the weight and the maximum (we define the weight of each edge to represent the communication demand between node pairs) in the undirected graph, so that it contains the largest number of edges, weight and maximum, and each node appears only once. it can preferentially establish wavelength connection for node pairs occupying most of the current network resources. This method of constructing wavelength

Copyright © 2020 Qishuang Zhu *et al.*
doi:10.18282/jnt.v2i1.836
This is an open-access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

connection can give full play to the spirit of optical network in logical topology construction Activity, so that the flow of large data flows through a hop path to communicate as much as possible. However, when the traffic is regional, it can be forwarded by nodes in the region in a short distance. This method of constructing wavelength connection according to global traffic causes wavelength connection must be wasted. as shown in Figure 1 (a) and (b), since the traffic is regional, the wavelength connection established according to b-matching is also regional. (a) indicates the wavelength connection established based on global traffic. (b) indicates the wavelength connection established according to regional traffic. In (a), the communication between c-d can be forwarded through c-a-b-d, so the wavelength connection between c-d can be deleted. similarly, the wavelength connection between g-h can also be deleted. Therefore, this paper proposes a new topology reconstruction mechanism for data center optical interconnection networks, which can be deployed on any optical network capable of reconstructing topology to match traffic patterns.

The mechanism includes a global topology generation module based on b-matching, a traffic identification module based on machine learning, and a topology clipping module based on maximizing the utilization rate of wavelength connection. Simulation results show that the mechanism proposed in this paper can give full play to the flexibility of optical networks, save bandwidth resources and improve the utilization rate of wavelength connection bandwidth without sacrificing network performance.

1. Topological reconstruction mechanism

In order to adapt to different traffic types, make full use of network bandwidth resources and realize more flexible reconstruction schemes, this paper proposes is a data center optical interconnection network topology reconstruction mechanism, as shown in Figure 1 (c), with the specific process as follows:

Step 1: Data Center Network Traffic Monitor Regularly Monitors Network Traffic to Obtain Traffic Demand Matrix TM;

Step 2: calculates the global topology with b-matching according to the traffic demand matrix TM;

Step 3: simultaneously sends the monitored TM to the traffic identification module based on machine learning to identify whether it is global traffic or regional traffic;

Step 4: cuts out the topology generated in Step2 in the topology cutting module based on the maximum utilization rate of wavelength connection according to the traffic identification result, and the output topology is the logical topology of the network. In order to realize this topology reconfiguration mechanism, we need to develop two modules in the network controller, which are traffic identification module and topology clipping module respectively. The traffic identification module is used to identify whether it is global traffic or regional traffic. Regional traffic identification actually identifies the number and size of regions and the nodes they contain. We can know the traffic directly from the information obtained by the application mode. But this may add additional development costs.

Therefore, based on machine learning, this paper uses the method[10] combining spectral clustering and convolutional neural network (CNN), the regional traffic is generated from k concurrent computing jobs, each job runs independently on a group of cluster nodes, so the relevant strong connections are all in one job. In the flow pattern recognition module. In the block, identifying the traffic pattern is to identify the number of jobs k in the traffic matrix first, and then cluster the nodes into k groups. The identification of the number k of industries can be regarded as a classification problem. Traffic matrix has the characteristics of graph. Classifying a traffic matrix can be viewed.

When the traffic pattern is identified, the identified traffic is sent to the topology clipping module based on maximizing the utilization rate of wavelength connection. If the traffic type is global traffic, this module directly outputs the global topology generated based on b-matching as the logical topology of the network. If the traffic type is regional traffic, the wavelength connection established by b-matching will be cut off. The number of hops between nodes in the region is small. In most cases, it can be forwarded by neighboring nodes to reach the destination node, avoiding the problems of low wavelength utilization rate and waste of bandwidth resources. Therefore, we will discuss how many wavelength connections are allocated to different regions by taking each region as a unit. The allocation principle is: the OCS edges established by b-matching will be sorted according to the weight from top to bottom, and the OCS edges will be deleted from bottom to top in turn. At the same time, we will observe the throughput change caused by deleting edges. We will define an effective value for each edge. When deleting this edge, the change caused by throughput

is very small. The utility value of this edge is small, which indicates that the traffic carried on these edges can be forwarded through other nodes. These wavelength connections do not contribute much to throughput, so these edges can be deleted. When deleting this edge causes great changes in throughput, the utility value of this edge is larger, which indicates that these edges carry a large amount of traffic. If the network is forwarded through other nodes, it will cause network congestion, and the total traffic that can be transmitted by the network per unit time will decrease, so these edges cannot be clipped. To sum up, the topology clipping module needs to evaluate the wavelength connections in each region separately, clip those edges with smaller utility values, and summarize them to obtain the final logical topology of the network.

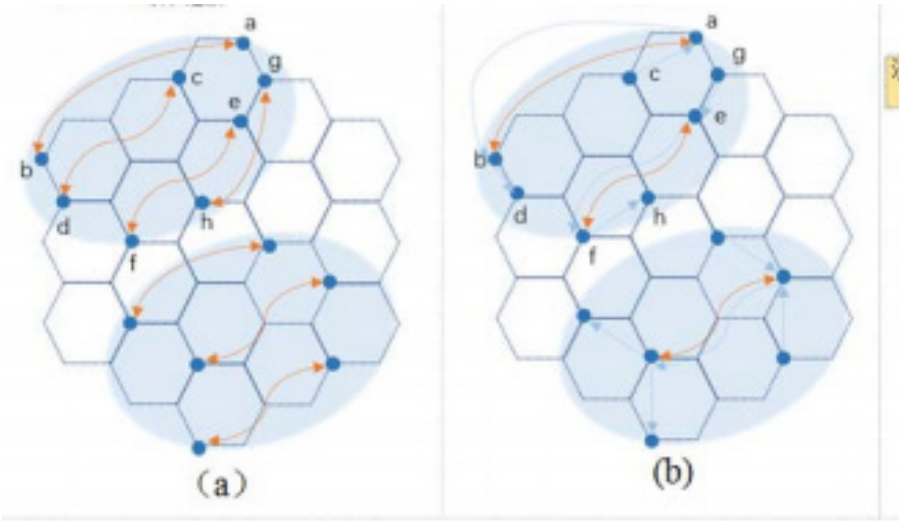
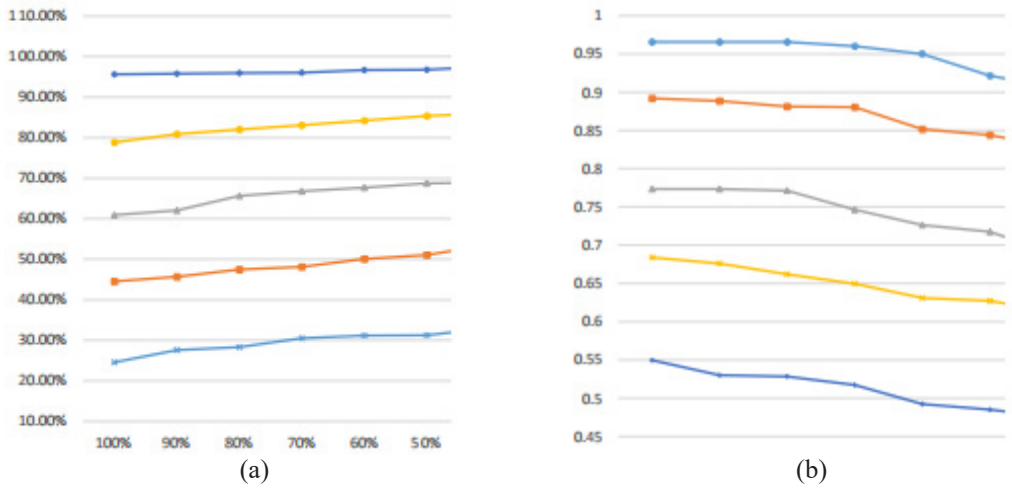


Figure 1. Wavelength connection comparison based on global traffic and regional traffic and flow chart of topology reconfiguration iguration mechanism. (a) wavelength connection based on global traffic using b-matching; (b) wavelength connection based on regional traffic.

2. Simulation analysis

In the simulation, we select the OpenScale[11] optical interconnection network to evaluate the topology reconstruction mechanism based on traffic identification. The OpenScale network uses optical fibers to interconnect top of rack (TOR: Top of Rack) switches into a network topology of hexagonal honeycomb structures. Each TOR switch also integrates an optical switching module, collectively referred to as an optical top of rack (OTOR: Optical Top of Rack), and each cellular unit communicates in an optical burst switching ring network. Each OToR module also has the function of optical add/drop multiplexer, which can support the dynamic establishment or removal of OCS connection between any two racks. In this paper, the OpenScale network is selected for evaluation, and any other network capable of topology reconstruction can also be selected for further evaluation.



(a) (b)

Figure 2. Changes in utilization rate of wavelength connections and throughput assigned to different OCS

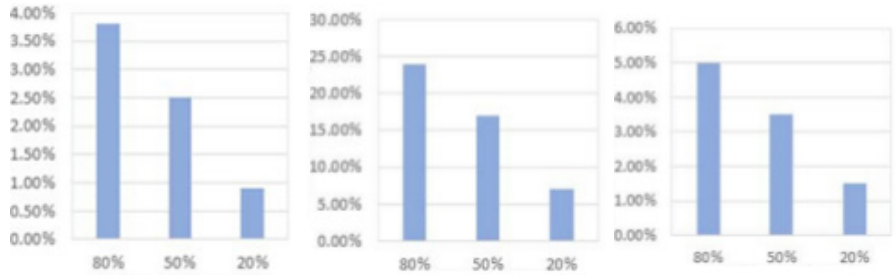
connections. (a) changes in utilization rate of wavelength connections assigned to different OCS connections; (b) changes in throughput assigned to different OCS connections.

In the topology clipping module, we need to allocate different OCS connections for different area sizes and traffic loads. In the simulation of this paper, we set the size of the area to be between 40-60 nodes to generate random integers according to uniform distribution. This paper takes 54 nodes as an example. We use the on/off data source[12] with Pareto distribution to generate traffic with self-similar characteristics. By changing α (representing the tailing degree of the Pareto distribution function) in the probability density function, we can set the load of the traffic matrix and allocate wavelength connections by discussing the throughput and wavelength connection utilization rate of different reconstruction schemes under different loads. This paper only considers the case of 1 degree of freedom (only one wavelength connection is allowed per node), then OCS can be built with b-matching with 27 sides. In the process of solving throughput and wavelength connection utilization, we all adopt the shortest path routing scheme.

The simulation results are shown in Figure 2. The abscissa indicates the allocation rate of OCS connections. When the traffic matrix load is 0.1, there is no significant difference in network throughput from allocating all OCS connections to only the 60% before allocation. When fewer OCS connections are allocated, the throughput drops significantly. At the same time, when 60% of OCS connections are allocated, Wavelength utilization rate increases by about 5%, so when traffic load is 0.1, we can allocate the top 60% of OCS connections generated by b-matching to the network, which can save 40% OCS connections without sacrificing network throughput and increase wavelength connection utilization rate by 5%; Similarly, when the traffic load is 0.2, the first 70% of the OCS connection is allocated, and the wavelength utilization rate is increased by about 5%; When the traffic load is 0.3, the OCS connection of the 80% before allocation will increase the wavelength utilization by about 5%; When the traffic load is 0.4, the throughput of the OCS connection with 20% dropped obviously. At this time, 90% of the OCS connection needs to be reserved, and the wavelength utilization rate increases by about 2%; When the traffic load is 0.5, the wavelength utilization rate is close to 100%, the network congestion is serious, and the logical topology reconstruction can no longer meet the network demand at this time. Therefore, when the traffic load is 0.5 or above, we can only consider increasing the degree of freedom to improve the network performance, which we will not discuss in this article.

In real traffic, the probabilities of global traffic and regional traffic are uncertain. In order to evaluate this topology reconstruction mechanism, we simulate in the following three cases respectively: large probability of regional traffic (global traffic appears with 20% probability, regional traffic appears with 80% probability), large probability of global traffic (global traffic appears with 80% probability, and regional traffic appears with 20%). The global traffic and the regional traffic all occur (the global traffic and the regional traffic each occur with a probability of 50%). in these three cases, the throughput decline rate, OCS connection saving rate and wavelength connection utilization improvement rate of the topology reconstruction mechanism proposed in this paper are analyzed respectively compared with those of the case without such a mechanism. the simulation results are shown in Figure 3. it can be seen that the topology reconstruction mechanism based on traffic identification can be small regardless of the probability of the regional traffic occurring.

The efficiency of the proposed mechanism is proved by saving the number of OCS connections while improving the utilization rate of wavelength connections at the expense of throughput.



(a) (b) (c)

Figure 3. The topology reconstruction mechanism based on traffic identification has different probability of occurrence of regional traffic. (a) different probability of occurrence of regional traffic and different throughput decline

rate; (b) different probability of occurrence of regional traffic and different OCS connection saving rate; (c) When the probability of regional traffic is different.

3. Conclusion

In order to give full play to the flexibility of data center optical network topology, this paper proposes a topology reconfiguration machine system based on traffic identification. The simulation results show that this mechanism can save network bandwidth resources and improve the utilization rate of wavelength connections on the basis of using b-matching to establish logical topology. Although this paper evaluates on a specific network, this reconfiguration mechanism can be applied to any data center optical switching architecture that performs topology reconfiguration.

References

1. Farrington N, Porter G, Radhakrishnan S, et al. Helios: A Hybrid Electrical/Optical Switch Architecture for Modular Data Centers[C]//Acm Sigcomm Conference. ACM, 2010.
2. Chen K, Singla A, Singh A, et al. OSA: An Optical Switching Architecture for Data Center Networks With Unprecedented Flexibility[J]. IEEE/ACM Transactions on Networking, 2018, 22(2):498-511.
3. Wang G, Andersen DG, Kaminsky M, et al. c-Through: part-time optics in data centers[J]. Acm Sigcomm Computer Communication Review, 2010, 40(4):327-338.
4. Zhang D, Guo H, Wu J, et al. A deterministic small-world topology based optical switching network architecture for data centers[C]//European Conference on Optical Communication. IEEE, 2014.
5. Kandula S, Sengupta S, Greenberg AG, et al. The nature of data center traffic: measurements & analysis[C]//Acm Sigcomm Conference on Internet Measurement Conference. ACM, 2009.
6. Roy A, Zeng H, Bagga J, et al. Inside the Social Network's (Datacenter) Network[C]//Acm Conference on Special Interest Group on Data Communication. ACM, 2015.
7. Xia Y, Sun XS, Dzinamarira S, et al. A Tale of Two Topologies: Exploring Convertible Data Center Network Architectures with Flat-tree[C]//Conference of the Acm Special Interest Group. ACM, 2017.
8. Xiaoqiao Meng LZ. Improving the scalability of data center networks with traffic-aware virtual machine placement[J]. Proceedings - IEEE INFOCOM, 2010, 54(1):1-9.
9. Osiakwan CNK, Akl SG. The maximum weight perfect matching problem for complete weighted graphs is in PC[C]//IEEE Second Symposium on Parallel & Distributed Processing. 1990.
10. Cen Wang, Hongxiang Guo, Xiong Gao, et al. Machine Learning Based Traffic Pattern Aware Topology Reconstruction to Optimize Application Performance in Optical DCNs[C]//OFC. IEEE, 2018.
11. Zhang D, Guo H, Chen G, et al. Analysis and experimental demonstration of an optical switching enabled scalable data center network architecture[J]. Optical Switching and Networking, 2016:S1573427716300042.
12. Walter W, Vern P, Taqqu Murad S. Self-similary and heavy tails: structural modeling of network traffic[J].Preprint, 1996: 27-53.

About PiscoMed Publishing

PiscoMed Publishing Pte Ltd is an international company established in 2011, setting up its headquarter office in Singapore.

PiscoMed Publishing started off with a focus in advancing medical research, however with the advancement of all areas of science, technology and medicines, PiscoMed have decided to venture into all areas of research, publishing quality journals that will support the scholarly and professional community across the globe. Our aim is to transform research into tangible findings that will form the basis for further research, guidelines, knowledge and more.

PiscoMed adopts the Open-Access approach of publishing which allow immediate visibility and access of research outputs and free usage of researchers' findings and results.

At present, the publisher has established more than 100 international academic journals and books in the fields such as medicine, engineering, education, finance, and agriculture. And the journals are indexed in academic databases such as Google Scholar, Europub, ResearchBib, PKP Index, DRJI, Semantic Scholar, CNKI, CQVIP. PiscoMed has academic publishing cooperation with internationally well-known universities and research institutions.

Copyright Notice

Authors submitting to PiscoMed journals agree to publish their manuscript under the Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0) where authors agree to allow third parties to share their work (copy, distribute, transmit) and to adapt it, under the condition that the authors are given credit, and that in the event of reuse or distribution, the terms of this license are made clear.

Authors retain copyright of their work, with first publication rights (online and print) granted to PiscoMed Publishing or the owner of the journal in question.

Contact Information

Address: 73 Upper Paya Lebar Road #07-02B-11 Centro Bianco, Singapore

Phone: +65 8822 5925

E-mail: contact@piscomed.com

ISSN 2661-4065

Pisco Med Publishing

PiscoMed Publishing Pte. Ltd

Address: 73 Upper Paya Lebar Road #07-02B-11 Centro Bianco Singapore 534818

Website: www.piscomed.com

Email: contact@piscomed.com