PPAdroid: An Approach to Android Privacy Protocol Analysis

Yongming Yao[1,2], Yulin Wang[2], Weiyi Jiang[2], Ziyuan Wang[3], Song Huang[1*]

[1] Command and Control Engineering College, Peoples Liberation Army Engineering University, China
[2] Tongda College, Nanjing University of Posts and Telecommunications, China
[3] School of Computer Science and Technology, Nanjing University of Posts and Telecommunications, China
yaoym@njupt.edu.cn, 1057468754@qq.com, 599039615@qq.com, wangziyuan@njupt.edu, hs0317@163.com

Abstract

With the continuous growth of the number of mobile applications, users may provide personal information to applications consciously or unconsciously. Privacy protocol could help users understand the privacy behavior of applications. However, users usually ignore the content of the privacy protocol due to the length of their context. To solve such a problem, we conduct an empirical study on privacy protocols of Android applications and propose a method to detect sentences related to personal information operations in privacy protocol documents. In our proposed method, a verb list, a negation list, and a noun list related to user information operations are created and utilized to detect sentences related to user personal information by the Stanford CoreNLP technique. The experimental results show that our proposed method is better than state-of-art keyword-based methods. Furthermore, our proposed method can help users understand the contents of privacy protocol documents in a limited way.

Keywords: Privacy protocol, Natural language processing, Android automation, Application market

1 Introduction

In recent years, with the continuous growth of the number of mobile applications, consumers have faced a variety of privacy violations. Smartphones could transmit rich data to app developers so users of mobile devices may be vulnerable to privacy violations and privacy abuses by numerous entities, app developers, analytics services, and AD networks [2]. These entities have unlimited access to sensitive information, including users' locations, contacts, identities, messages, and photos. Both at home and abroad, user privacy protection has been strengthened in recent years, and the privacy security of mobile applications has gradually become the focus.

Such a risk is extremely important due to the large number of smartphone users. The iiMedia Research reported that the number of active users of third-party mobile app stores in China reached 472 million in 2018 and is expected to reach 500 million in 2021 [3]. The market of third-party mobile app stores tends to be stable. Users expect mobile app stores to improve security, richness, and recommendation accuracy. China's National Bureau of Statistics reported that there are 986 million mobile Internet users in China, making it one of the world's largest mobile users, led by Android users [4]. In addition, due to the restrictions of the Google Play Store in China, some other Android app markets were born, such as Huawei AppGallery [5], Mi App Store [6], Anzhi App Store [7], Tencent App Store [8], Wandoujia [9], etc.

Unfortunately, according to the research and analysis of the Domestic Android market and Foreign Google market by Wang et al. [10], potential malicious applications are more common in China's applications market than in Google Play. In addition, according to the Application Notice of China National Mobile Internet Application Security Management Center [11], it could be found that the most illegal applications are caused by non-compliance with privacy protocols.

To protect privacy of users, a privacy policy document should be published when uploading applications [12]. In such a document, the developers are required to declare how users' privacy information is used, collected, or shared to make users know how the privacy information is used and to better protect their privacy. Although the developers of apps publish privacy policies to inform users of the potential privacy risks of their software, it is difficult for common users to intuitively determine the correctness of the documentation due to the length of their context. In addition, they are unaware when apps collect or share information beyond what is stated in the privacy policy. Related studies [13]-[16] have shown that the inconsistency of many application behaviors with their privacy regulations carries the risk of misuse of permissions and privacy breaches. According to the research of Alani et al. [17], among 4027 Android users interviewed, 68.61% of them are not aware of which applications will access their private data. Users either do not read the permissions required by the application or directly ignore it.

Although there have been many works on Android application privacy protocols, these works still have many limitations. (1) the current work mainly focuses on foreign Google play market applications; (2) the application privacy protocols in the existing studies are all in English; (3) the Chinese laws and regulations on personal privacy are imperfect. The above situation indicates a gap between the current work and the actual situation of Android applications in China, so there is a need to study the privacy protocols of Android applications in the Chinese application market. Based on the analysis of privacy protocols of Android applications in China's application markets, this paper proposes a method to detect sentences related to personal information operations in privacy protocol documents. The experimental results outperform the keyword approach and can help users read the privacy protocol documents in a limited way.

*Corresponding Author: Song Huang; E-mail: hs0317@163.com
DOI: 10.53106/160792642022052303014

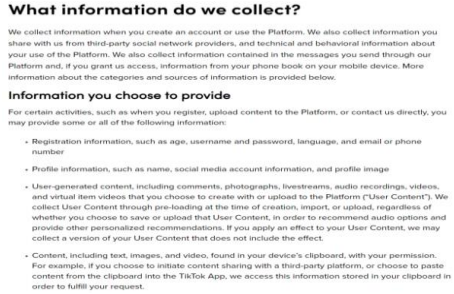
A preliminary version of the partial context of this paper was presented at the 8th International Conference on Dependable Systems and Their Applications (DSA) [1]. Its contributions included (1) We created a dataset from the Huawei AppGallery in China, including 854 privacy protocol texts. (2) We propose a method to detect the operation sentences related to a user's personal information in privacy protocol documents, which combines verbs, nouns, and negatives. (3) In 200 application privacy protocol documents (48064 sentences in total), our evaluation shows that our method has high accuracy in detecting the sentences related to personal information operation in privacy protocol documents. This extended version provides some expanded contributions including the following: (4) We investigated the market share of third-party Android app stores in China, and selected Huawei AppGallery, which had a high proportion and few official reactions to survey conducted a survey on the application privacy protocol. It was found that 10% of the privacy protocol still had many defects and errors. (5) We use our method to experiment with the privacy protocol of 10 categories of applications, and have good results in the analysis of the privacy protocol of each category of applications.

The rest of this paper is structured as follows. Section 2 introduces the background of Android privacy protocol, Android automation, and natural language processing technology. Section 3 discusses motivation and privacy protocols in the app store for empirical analysis. Section 4 introduces the design and implementation of our method and analyzes the sentences in the privacy protocol document. Section 5 evaluates our method. Section 6 discusses validity threat, Section 7 discusses the related work, and Section 8 summarizes and discusses future work.

2 Background

2.1 Android Privacy Protocol

The information types included in the Android Privacy protocol mainly include the information that can reflect the user's identity, the personal information that may be used or shared when the application is running, and the purpose of information use.



What information do we collect?

We collect information when you create an account or use the Platform. We also collect information you share with us from third-party social network providers, and technical and behavioral information about your use of the Platform. We also collect information contained in the messages you send through our Platform and, if you grant us access, information from your phone book on your mobile device. More information about the categories and sources of information is provided below.

Information you choose to provide

For certain activities, such as when you register, upload content to the Platform, or contact us directly, you may provide some or all of the following information:

- Registration information, such as age, username and password, language, and email or phone number
- Profile information, such as name, social media account information, and profile image
- User-generated content, including comments, photographs, livestreams, audio recordings, videos, and virtual item videos that you choose to create with or upload to the Platform ("User Content"). We collect User Content through pre-loading at the time of creation, import, or upload, regardless of whether you choose to save or upload that User Content, in order to recommend audio options and provide other personalized recommendations. If you apply an effect to your User Content, we may collect a version of your User Content that does not include the effect.
- Content, including text, images, and video, found in your device's clipboard, with your permission. For example, if you choose to initiate content sharing with a third-party platform, or choose to paste content from the clipboard into the TikTok App, we access this information stored in your clipboard in order to fulfill your request.

Figure 1. Privacy protocol of Douyin

For example, Figure 1 shows part of the privacy protocol of Douyin [18], it is called TikTok in the Google Play Store, which is the most famous short video application in China. It is very famous in China and has a large number of downloads and users. Examples show how the application uses information in statements.

2.2 Android Automation

This paper will obtain the privacy protocol documents in the application market through Android automation technologies and use them as data set. The following describes the relevant techniques of Android automation:

Appium: Appium is an automated testing framework that can test various applications in Android and IOS terminals. It supports a variety of development languages (Python, Java, etc.). On the Android side, appium realizes the automatic test of the app by calling the command of UI Automator based on the webdriver protocol [19].

AAPT: Android resource packaging tool. Use the command "AAPT dump bagging + APK absolute path" to obtain the APK related information, including appPackage and appActivity [20].

UI Automator: We can get a screenshot of the current interface and obtain the structure and control information. We can locate the control according to the relevant information such as resource ID, class, text, or XPath to simulate the user's operation [21].

2.3 Stanford CoreNLP

NLP technology refers to the technology of analyzing and understanding human language to understand and extract information from text. NLP technology can deal with a large number of text data and solve various problems [22]. This paper uses the Stanford CoreNLP tool, which is a natural language processing toolkit integrated with many practical functions [23]. We will briefly introduce the key NLP technologies used in this work:

- **Tokenization:** In other words, the sentence will be split into words. Using the word_tokenize method in Stanford CoreNLP can determine the function of word segmentation.
- **Part-of-Speech Tagging (POS):** That is, using pos_tag for a given sentence method can label the part of speech of the word after segmentation. For example, VV stands for verb, NN stands for noun, and AD stands for adverb.

3 Motivation

3.1 An Empirical Study of Privacy Protocols in China's Application Markets

A recent report by App Annie's [24] shows that this has spurred the rapid growth of the mobile app market due to soaring demand for video conferencing, food delivery, grocery shopping, online learning, video streaming, and telemedicine. Because of COVID-19, mobile shopping and mobile payment have gradually become indispensable applications in users' cell phones. Meanwhile, there are about 900 million smartphone users in China, 80% of whom use the Android operating system, and every user has used more than one mobile application market. Based on these data, it can be seen that it is necessary to study the mobile application market in China. Figure 2 shows the share of active users in China's Android App Store in July 2020. You can see the active users of the main app markets in China. Unlike Google Play, you can see that there are many Android application markets in

China, each of which has many active users and they are from different companies; each of them has many app categories with wide coverage, so it is reliable to choose them for the study. In this section, we will investigate the privacy protocols of these 10 application markets, and the specific investigation methods and conditions will be explained in the subsequent sections.

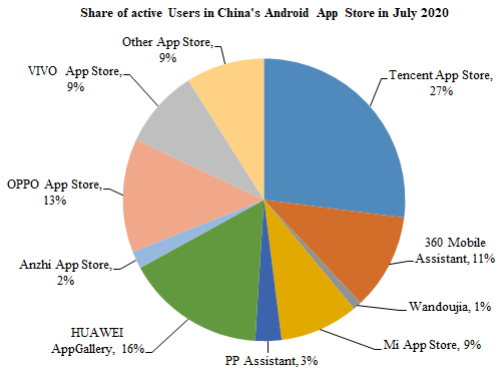


Figure 2. Share of active users in China's Android app store in July 2020

Since the privacy protocol will be available later, some concepts will be explained below.

1) Privacy protocol link. Enter the link in the web page to jump to the privacy protocol page.

2) Privacy protocol document. Save the content of the page that jumps to the privacy protocol link into the text.

3) Correct privacy protocol link. You can open the privacy protocol link correctly and go to the correct privacy protocol website.

4) Invalid privacy protocol link. The privacy protocol link does not open, or the redirect screen is not the privacy protocol website, but the official website of the application.

5) No privacy protocol link. The link to the application's privacy protocol was not successfully obtained through automated means, and the application's download page does not provide a privacy protocol.

6) Valid privacy protocol document. The privacy protocol file obtained through the privacy protocol link is not empty and the content is correct and unique.

7) Invalid privacy protocol document. There are two cases: The first is that the privacy protocol file is in English. The second one is that the content of the privacy protocol file is empty.

We studied the above 10 application marketplaces. Our research includes four aspects.

1) Privacy protocol link detection in the Android application market. Investigate whether web and mobile applications in the Android application market have privacy protocol links.

2) Identifying the Android application market to be investigated. By analyzing the data of the Android application market, the final Android application market to be investigated is determined from several dimensions.

3) Acquisition of Android application privacy protocol documents. Obtain the privacy protocol link from the webpage or mobile terminal of the application market, and further obtain the corresponding privacy protocol document.

4) Analysis of the research results of Android application privacy protocols. The results of the privacy protocol links and privacy protocol documents of the application market are analyzed.

3.2 Analysis of The Problems in China's Application Markets

Firstly, we conduct a preliminary screening based on the active share of users in the application market. We select the top three APP application markets with market share. They are Application Treasure, HUAWEI AppGallery, and Mi App Store. Then, we discuss according to two aspects to determine the application market under investigation.

On one hand, after investigating the privacy protocol link between the app market's web and mobile terminals, Application Treasure's website does not have a privacy protocol link, and both the Mi App Store and HUAWEI AppGallery's privacy protocol links exist; on the mobile terminal, there is no privacy protocol link in the Mi App Store. Both the App Store and HUAWEI AppGallery have a privacy protocol link. In this respect, HUAWEI AppGallery is superior to the Mi App Store and App Store in this regard.

Table 1. Statistical table of problems involved in the application

The issues involved	Applications	Percentage
Collection of personal information in violation of regulations	145	57.3%
App compulsory, frequent, and excessive requests for permissions	45	17.8%
Illegal use of personal information	26	10.3%
Collect personal information beyond the scope	15	5.9%
Force users to use the targeted push function	9	3.6%
App information on the app distribution platform is clearly not in place	6	2.4%
To deceive and mislead users to download the app	6	2.4%
To deceive and mislead users to provide personal information	1	0.4%
Total	253	100.0%

On the other hand, in the announcement issued by the Ministry of Industry and Information Technology on January 22, 2021, after 10 tests, there were 157 problematic applications. Among them, the app accounted for 22.3% in Application Treasure, 12.0% in the Mi App Store, and 8.8% in the HUAWEI AppGallery. Application Treasure is the application market with the largest market share. The proportion of problematic apps is almost three times that of the HUAWEI AppGallery and twice that of the Mi App Store. The proportion of apps with problems in the Mi App Store is also slightly higher than that of the HUAWEI AppGallery. Huawei's application market has the least problems. On the whole, the apps of HUAWEI AppGallery are better than those provided by Application Treasure and Mi App Store. According to the above notice from the Ministry of Industry and Information Technology, there are many suspected problems with illegal applications. We have made a statistical table of the problems involved for several specific problems,

as shown in Table 1. There are two or even three problems with one application at the same time. In the case of 157 apps, there were a total of 253 problems, of which 73.9% were illegally collected, used personal information, collected personal information beyond the scope, and misled users to provide personal information. Questions accounted for 17.8%. It can be seen that most of the questions are related to users' personal information, and the details related to users' personal information should be mentioned in the privacy protocol. We have introduced relevant information on this point above. By the regulations of the "Methods for the Determination of the Collection and Use of Personal Information in Violations of App Laws and Regulations", the privacy protocol shall explain how to collect, use, and disclose user information.

3.3 Analysis of Android Application Privacy Protocols of Android App Store in China

According to our preliminary research on the domestic mainstream application market, we found that in the 10 application markets with the most active users, there are the following problems: the application market does not have a web page and neither the web page nor the mobile site has a privacy protocol link or a sensitive permission list. To this end, we research the privacy protocol situation in the domestic application market to determine which application market's privacy protocol will be used as the subsequent test data set for analysis. The problems in the market were analyzed in two aspects, and the third-party application market was finally determined to be the HUAWEI AppGallery.

After determining the source of the data, we proceeded to obtain relevant information in the HUAWEI AppGallery. Although the web version of HUAWEI AppGallery provides a privacy protocol, it cannot directly use the crawler to obtain the corresponding link. Therefore, we consider obtaining the privacy protocol from the mobile APP. The method we obtain has been explained in detail in the previous article. We selected applications from ten categories in the application market and each category crawled about 110 privacy protocol links. After manual verification, we concluded that the following situations existed: 1) Correct privacy protocol Link; 2) Invalid privacy protocol link; 3) Unprovided privacy protocol link; 4) Effective privacy protocol document;5) Invalid privacy protocol document.

Among them, we classify the English version of the privacy protocol document into the invalid privacy protocol document. From two aspects, firstly, we are investigating the domestic application market. The English privacy protocol should not have appeared, and the English version of the privacy protocol greatly increases the difficulty for users to read. Secondly, we aim to study domestic Chinese privacy protocols, and English will cause errors in our follow-up progress. Combining these two factors, we classify the English privacy protocol as an invalid privacy protocol document. According to the obtained 1,069 privacy protocol links, we conducted a preliminary analysis of the results and only analyzed whether the privacy protocol links were correct or not. The analysis is shown in the Table 2. It can be seen that invalid privacy protocol links accounted for 9.07%, which means that 9 application download interfaces in every 100 applications provide incorrect privacy protocol links. In other words, they cannot open or jump to the non-privacy protocol document interface. Moreover, there are 1.87% of applications that do not provide a link to a privacy protocol.

Table 2. Privacy protocol link analysis

Application category	Links to privacy protocol documents	Correct privacy protocol links	Invalid privacy protocol links	Privacy protocol links not provided
Media entertainment	105	99	5	1
Tools	110	98	10	2
Social communication	106	104	2	0
Education	110	102	7	1
News reading	110	103	4	3
Beautification	110	85	20	5
Delicacy	106	88	16	2
Travel navigation	110	99	8	3
Tourist accommodation	100	86	11	3
Shopping comparison	102	88	14	0
Total	1069	952	97	20
Percentage	100%	89.06%	9.07%	1.87%

Finally, we manually checked and sorted the data. The data situation is shown in Table 3. The relevant meaning has been outlined at the beginning of this section. It can be observed from the data that the number of valid privacy documents accounted for 90%, and the invalid privacy protocol documents accounted for 10%. According to the notice from the Ministry of Industry and Information Technology mentioned above, Huawei AppGallery has problems accounting for 8.8%, but in our actual research, privacy protocol links have problems accounting for about 11%, and invalid privacy protocol documents account for 10%, which is higher than the announcement of the Ministry of Industry and Information Technology. It can be seen that there are still problems with the jump of the privacy protocol, and whether the content of the privacy protocol is correct is also a question worth considering.

Table 3. Analysis of privacy protocol documents

Application category	privacy protocol documents	Effective privacy protocol document	Invalid privacy protocol document
Media entertainment	87	82	5
Tools	100	90	10
Social communication	95	93	2
Education	98	91	7
News reading	96	92	4
Beautification	101	81	20
Delicacy	96	80	16
Travel navigation	99	91	8
Tourist accommodation	86	75	11
Shopping comparison	93	79	14
Total	951	854	97
Percentage	100%	90%	10%

4 Design of Detection Method

This section will introduce our overall framework, which is composed of four components: data acquisition, preprocessor, generate sentences object set, and detection module design. The overall framework is shown in Figure 3, we will explain the four components in detail below.

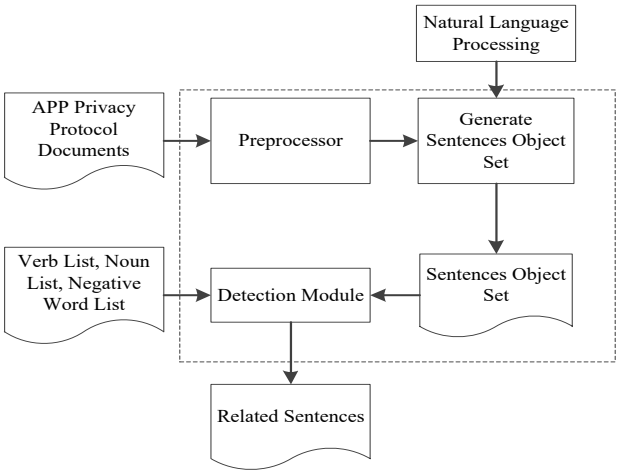


Figure 3. Overall framework of the design

Firstly, we obtain the data set and get the privacy protocol document as our input part; Secondly, preprocess the statements in the privacy protocol document; Then, on this basis, the statement object set is generated after natural language processing; Finally, the detection module is designed to detect and match verbs and nouns in the statement object collection, to output relevant statements.

4.1 Privacy Protocol Document Analysis

4.1.1 Analysis of The Components of Privacy Protocol

Most apps now have user login functions. The following sentence is a common one in privacy protocol "We collect such information (personal information) to complete the login of APP users. If you refuse to provide it, you may not be able to complete the login and then cannot use the app normally." We can see that the sentence mainly consists of five parts, including action executor, action, resource, purpose, and result.

Action executor: It is the main body to collect, save, or share user's personal information. It's usually considered a subject in a sentence, such as "we" in the above sentence.

Action: The actions performed by the action executor include collection, saving, sharing, etc., such as "collect" in the above sentence.

Resource: The object of action execution, namely the user's relevant privacy information or sensitive rights, such as "personal information" in the above sentence.

Purpose: The reason for collecting personal information or opening certain permission is to complete subsequent operations, such as "complete the login of APP users" in the above sentence.

Result: If the required action is not executed or the required permission is not opened, the developer will inform the user of the consequences caused by the absence of this action or permission. Such as "you may not be able to complete the login and then cannot use the app normally."

The sentence structure is shown in Figure 4.

We collect such information (personal information) to
Action executor Action Resource
complete the login of APP users. If you refuse to provide it, you may
Purpose
not be able to complete the login and then cannot use the app normally.
Result

Figure 4. Sentence structure

4.1.2 Classification of Action Verbs

According to the research results of Breaux, T. D. et al. [25], they divided the common verbs in privacy regulations into four categories, including collection, use, preservation, and transfer. We did not directly use their results. We used our analysis to combine with their results to make up for the possible omissions caused by the differences between China and the UK.

Firstly, we process our 854 privacy protocol document datasets through the part of speech tagging of Stanford CoreNLP, and get the part of speech of each word in each sentence, with a total of 808028 verbs. Then, we calculate the word frequency by using the counter class in the python collection module. We use the verbs with high frequency and combine the research results of Breaux, T.D. et al. [25] to make up for the possible omission because of the differences between Chinese and English. We continue the verb classification they use: Collect, Use, Retain, Share.

As shown in Table 4, we select common verbs as valid verbs and add them to the list of privacy protocol verbs to determine whether they are related to the operation of user information. In this process, we found that the negative meanings of words also need to be considered, such as: no collection, no disclosure, no sharing, etc. Users also need to know which sentences in the privacy protocol document do not collect information, so we created a list of negative words, and combined with the negative words to judge.

Table 4. Privacy protocol verblist

Type	Number	Similar Meanings
Collect	11	collect, gather, acquire, obtain, gain, supply, record, trace, need, provide, achieve
Use	8	access, include, handle, view, use, query, fetch, utilize
Retain	3	save, store, retain
Share	8	exchange, publish, share, show, sell, announce, exhibit, transfer

4.1.3 Identification of the Scope of Personal Information

The document on the scope of necessary personal information for common types of mobile Internet applications is a notice issued by China's State Internet Information Office. The regulations specify the scope of necessary personal information for common types of apps. For example, in the category of instant messaging, the primary function service is "to provide text, picture, voice, video and another network instant messaging services," the necessary personal information includes: 1. cell phone number of registered users.

2. account information: account number, instant messaging contact account list.

The provisions on the scope of necessary personal information for common types of mobile Internet applications" [26] clearly points out the necessary information to be collected by various types of app.

We use the scope specified in the regulation as our noun resource to associate with verbs and create a list of nouns.

4.2 Four Components of Detection Method

4.2.1 Data Sets Acquisition

We design the method to obtain the privacy protocol document, as shown in Figure 5. We explain the specific process as follows:

- After obtaining the HUAWEI AppGallery installation package, we use the AAPT instruction "AAPT dump bagging Huawei.APK" to get its appPackage "com.Huawei.Appmarket" and appActivity "com.Huawei.Appmarket.Mainactivity".
- Use ADB Android bridge to link an Android simulator or a real machine and use the corresponding content obtained by AAPT to set the appium parameter.
- Obtain the information of each control in the UI interface of the HUAWEI AppGallery through UI Automator and download the privacy policy and permission list in APP through the Selenium automation script.
- After obtaining the privacy protocol link, we use the requests module and re-module in Python to write a web crawler to obtain the privacy link content and use the regular expression to filter irrelevant characters, and finally get the documents in the privacy protocol link.

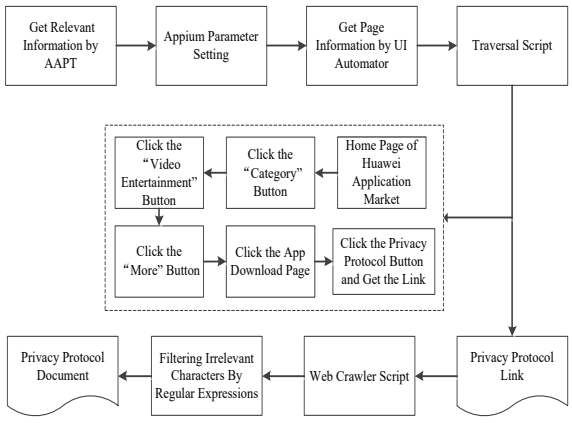


Figure 5. Data sets acquisition

4.2.2 Preprocessor

The preprocessor accepts the txt file of the privacy protocol document and performs the following preprocessing tasks:

Sentence Boundary Processing: Since a sentence may be split in the privacy protocol document, we first link all the sentences together to deal with the wrong segmentation. Secondly, the Chinese sentence format is usually ".", "!", "?"…". For the end, we use Python to read the entire txt text and use the replace method to replace the end character with the newline character to achieve the correct segmentation of all sentences.

Filtration Processing: There may be irrelevant characters in a document, which may lead to errors in subsequent analysis. Therefore, irrelevant characters need to be processed, such as "*" and "☆" and so on. However, due to too many forms of special symbols, it is hard to filter them directly. Therefore, we use regular expressions to formulate filtering rules, Filter out what we need, Chinese characters, punctuation, English letters, etc.

4.2.3 Generate Sentences Object Set

Take the preprocessed document as input, and the sentences are processed as follows:

- **Part-of-Speech Tagging(POS):** That is to say, describe the part of speech of words in the sentence, such as "We will collect your personal information." Some of the results are as follows: ('We', 'PN'), ('will', 'AD'), ('collect', 'VV'), ('information', 'NN'), where VV is verb, NN is noun, and PN is pronoun.
- **Generate Sentences Object Set：** That is, after each sentence of the privacy protocol document is processed, the output dictionary set contains the index value, word content, word start and end position, part of speech tagging, etc. During the experiment, the value obtained by the parse method of Stanford CoreNLP is of type string, which is difficult to be used as data input for subsequent processing. Therefore, we modified the method to make its output value of dictionary type, as shown in Figure 6.

```
{'index': 1, 'word': 'We', 'originalText': 'We', 'characterOffsetBegin': 0,
'characterOffsetEnd': 2, 'pos': 'PN'}
{'index': 2, 'word': 'will', 'originalText': 'will', 'characterOffsetBegin': 2,
'characterOffsetEnd': 3, 'pos': 'AD'}
{'index': 3, 'word': 'collect', 'originalText': 'collect', 'characterOffsetBegin': 3,
'characterOffsetEnd': 5, 'pos': 'VV'}
```

Figure 6. Sentences object set

4.2.4 Detection Module

We aim to judge whether the sentences in the privacy protocol document are related to the operation of the user's personal information according to the verb, noun, and negative word list in Section 3. We use the following methods to judge: first, we read the set of sentences objects from the privacy protocol document, and input the data in the form of a dictionary; Secondly, the POS field (part of speech label) is matched to judge whether it is a verb, and the word field is used to judge whether it is a verb in the list(As shown in lines 1 to 3 of Algorithm 1); Then, if it is a verb in the list, it will traverse up and down according to the index value to find the modifier of the verb (such as "will", "no" and so on) and the noun associated with the verb; If the modifier is negative and the noun exists in the noun list, the sentence is recorded as "negative information correlation"; If the modifier is a non-negative word and the noun exists in the noun list, the sentence is recorded as "positive information correlation"; If the nouns do not match, the sentence is marked as "irrelevant operation sentence" (As shown in lines 4 to 14 of Algorithm 1); Finally, a new document marked as "positive information correlation" and "negative information correlation" is output (As shown in lines 15 of Algorithm 1). As shown in Algorithm 1, this is our code fragment to realize this function.

Algorithm 1. Sentence matching

```
Input: sentence, flag, dict
Output: sentence
1:flag=false
2:pos= Dict['pos']
3:word=Dict['word']
4:while (pos=='VV') do
5:   if word in list_dong then
6:    if flag==true then
7:       break
8:    else
9:       if word_2=='NN' and word_2 in list_ming then
10:       if word_3=='AD' and word_3 in list_dong then
11:          flag=true
12:          sentence=sentence+'positive information correlation'
13:       else
14:          flag=true
15:          sentence=sentence+'negative information correlation'
16:       end if
17:      end if
18:   end if
19:   end if
20:end while
```

5 Evaluation

This section will evaluate our method. We aim to correctly judge the sentences related to personal information operation in privacy protocol documents to help users see the relevant sentences more intuitively and decide whether to install this application.

We will answer the following research questions:

• **RQ1:** What are the precision, recall, F-score, and accuracy of our method in identifying sentences related to user information operations?

• **RQ2:** How effective is our method in identifying sentences related to user information operations, compared to keywords-based retrieval?

•**RQ3:** How many types of personal information problems can be detected by our proposed method?

5.1 Data Set

According to the report of iiMedia Research, we obtained the mainstream app market in China and conducted research and analysis on it. Combined with the relevant reports of the Ministry of industry and information technology, we finally determined the source of our data set from HUAWEI AppGallery. By eliminating the duplicate privacy protocol and English privacy protocol, we further reduce the data set. Finally, we retained 854 privacy protocol documents. We obtained 10 categories of applications, and randomly selected 20 privacy protocol documents from each category for the experiment. In general, 200 privacy protocol documents were analyzed with a total of 48064 sentences.

5.2 Evaluation Setup

We conduct manual analysis on the privacy protocol documents selected by us according to the "provisions" mentioned in Section 4.1 and "Information Security Technology—Personal Information Security Specification" [27] (GB/T 35273-2020) and mark the sentences related to the operation of personal information. When a sentence describes the operation of personal information in a positive tone, we mark it as "positive information correlation"; When a sentence describes the operation of personal information in a negative tone, we label it as "negative information correlation". When the sentence is irrelevant to personal information, we label it as an "irrelevant operation sentence". Among them, "positive information correlation" and "negative information correlation" are sentences related to the user operation, which are collectively referred to as "information related sentences". The reason why we make a distinction is to let users see more intuitively which sentences mean to collect information and which sentences mean not to collect information.

According to the results calculated by our method and the results of manual annotation, we can get the values of true positive, false positive, true negative, and false negative. The meanings are explained as follows: 1) true positive (TP)：In this paper, the manual annotation and the experimental results are "positive information correlation" and "negative information correlation" at the same time. 2) false positive (FP): In this paper, it is manually labeled as "irrelevant operation sentences", while the experimental results are "positive information correlation" or "negative information correlation". 3) true negative (TN): In this paper, both the manual annotation and the experimental results are "irrelevant operation sentences". 4) false negative (FN): In this paper, it is manually labeled as "positive information correlation" or "negative information correlation", while the experimental calculation result is "irrelevant operation sentence". Table 5 shows our results.

Table 5. Statistical results

Type	E	N
Positive Information Correlation	14204	15108
Negative Information Correlation	962	1012
Irrelevant Operation Sentences	32898	31944
Total Number of Sentences	48064	48064

#E: Results of experiment calculation #N: Results of manual annotation

In our research field, we often have the following evaluation indicators: precision, recall, F-score, and accuracy.

The precision rate refers to the ratio of the number of true positives (TP) to the number calculated (TP + FP):

$$\text{Precision}=\frac{TP}{TP+FP} \quad (1)$$

The recall rate refers to the number of true positives (TP) accounting for the actual number of manual verification (TP + FN):

$$\text{Recall}=\frac{\text{TP}}{\text{TP+FN}} \quad (2)$$

The F-score refers to the weighted harmonic mean of precision and recall:

$$\text{F-score}=\frac{2*\text{Precision}*\text{Recall}}{\text{Precision+Recall}} \quad (3)$$

The accuracy rate refers to the proportion of the sum of true positive and true negative in the total number:

$$\text{Accuracy}=\frac{\text{TP+TN}}{\text{TP+FP+TN+FN}} \quad (4)$$

5.3 Experimental Results and Analysis

In this section, we will answer **RQ1** and **RQ2**.

Answer1： For **RQ1**, we calculate the evaluation index to evaluate the effectiveness of our method in identifying the user's information-related operation sentences in privacy protocol documents. Our results show in Table 6.

The first column in Table 6 shows the application category**,** including Media entertainment, tools, social communication, education, etc. The other columns are explained as follows: T is the total number of sentences by our method, TP is true-positive, FP is false positive, TN is true-negative, FN is false-negative, P is precision rate, R is recall rate, F is F-score, and A is accuracy rate. As can be seen in Table 7, according to different types of APP privacy protocols, using our method for testing all achieved good results. For the entertainment class privacy protocol, we reached the highest accuracy. For the education class privacy protocol, our effect is slightly lower than other types. The results show that 92.0%, 86.6%, 89.2%, and 92.9% of the 48064 sentences in 200 privacy protocol documents can be achieved respectively.

Table 6. Comparison with manual verification

Application category	T	TP	FP	TN	FN	P	R	F	A
Media entertainment	1864	1756	108	5355	334	94.21%	84.02%	88.82%	94.15%
Tools	1311	1195	116	3122	178	91.15%	87.04%	89.05%	93.62%
Social communication	1996	1852	144	4797	282	92.79%	86.79%	89.69%	93.98%
Education	1515	1413	102	1845	261	93.27%	84.41%	88.62%	89.98%
News reading	1608	1495	113	4312	238	92.97%	86.27%	89.49%	94.30%
Beautification	1081	951	130	2198	86	87.97%	91.71%	89.80%	93.58%
Delicacy	1622	1495	127	2702	227	92.17%	86.82%	89.41%	92.22%
Travel navigation	984	916	68	1417	148	93.09%	86.09%	89.45%	91.53%
Tourist accommodation	1400	1247	153	2834	153	89.07%	89.07%	89.07%	93.02%
Shopping comparison	1785	1640	145	2156	253	91.88%	86.63%	89.18%	90.51%
Total	15166	13960	1206	30738	2160	92.05%	86.60%	89.24%	93.00%

Answer 2： For **RQ2**, we answer **RQ2** by comparing our method to a keyword-based retrieval approach in identifying the user's information-related operation sentences.

Table 7 shows the comparison between our method and keywords-based retrieval. The last line shows how far our method has improved with keyword-based retrieval. Among them "ΔP, ΔR, ΔF, and ΔA" are the percentage of improvement in precision rate, recall rate, F-score, and accuracy rate respectively, it shown that the accuracy, F-score, and accuracy increased by 44.6%, 25.2%, and 25.6% respectively, but the recall rate decreased by 11.8%.

Table 7. Comparison with keywords-based retrieval

Evaluation Index	P	R	F	A
Our Method	92.00%	86.60%	89.20%	92.90%
Keywords-Based Retrieval	47.40%	98.40%	64.00%	67.30%
Percentage of improvementΔ	44.60%	-11.80%	25.20%	25.60%

We next present illustrative examples of how our method performs better than keywords-based retrieval. For example, consider the sentence "please keep your WeChat account, mobile phone number, and verification code properly". The keywords-based retrieval method judged it as a sentence related to personal information operation. However, the verb "keep" is not an application program's operation on a user's personal information, Therefore, it can not be wrongly judged as a sentence related to personal information operation. Another example is "personal information refers to all kinds of information with...". This kind of sentence is to explain personal information, and it will be wrongly judged as a sentence related to personal information operation. However, our method can correctly classify this kind of statement as an irrelevant operation sentence.

Finally, we will discuss on why our method caused a decline in recall in comparison to keywords-based retrieval. The main reasons are as follows: first of all, the verbs are not perfect. There are a very small number of uncommon verbs that are not in our Verb List, so some sentences are not judged correctly; In addition, the sentence pattern mismatch makes us

miss some sentences related to personal information operation. One thing that needs to be explained is the lack of nouns. The noun list lacks a very small number of unconventional nouns, which leads to the omission. However, because the keywords-based retrieval method uses the same ranking list as our method, it is not the main reason for the decline of recall rate. To ensure the authenticity of our method, we did not modify the above problems. In the future, we should update the list of verbs and nouns.

Answer 3: Table 8 shows the types of personal information problems found by our method. Among the 200 application privacy protocols we collected, a total of 25 application privacy protocols were found to have some problems using the method we proposed, including 10 with Collection of personal information in violation of regulations, 5 privacy protocols with irregular use of personal information, 6 with the over-scope collection of personal information, and 4 with other problems.

Table 8. Types of personal information problems found by our method

The issues involved	Personal information problems found by our method
Collection of personal information in violation of regulations	10
Illegal use of personal information	5
Collect personal information beyond the scope	6
Other problems	4
Total	25

6 Threats to Validity

Threats to External Validity: It mainly includes the data sets used in our assessment. The data set we use is only from the data set provided in the HUAWEI AppGallery, and there is no other application market. Secondly, we only study the application privacy protocol documents in Chinese, but not English, and there are some limitations.

Threats to Internal Validity: It is mainly about the accuracy and completeness of our Verb List and noun list. To reduce the threat, our verbs come from the research of Breaux, T. D. et al. [25], and we calculate the word frequency of all the verbs in our sentences and screen out the high-frequency effective verbs to make up for the omission caused by the differences between Chinese and English; On the other hand, we determine our noun list according to the terms mentioned in the "provisions".

7 Relate Work

The proposed method involves some research fields, such as privacy protocol analysis, natural language processing, and Android Market Analysis. Next, we discuss the related work of the proposed method in these fields.

Yu L et al. [28] think that because the application code may be written by others in the form of outsourcing, or there is a third-party library that does not know the source code, the author who writes the privacy protocol may not understand the corresponding content, which leads to difficulties or errors in writing the privacy protocol; There have been similar studies (such as AppProfiler [29]) before, but AppProfiler is done manually. They proposed and developed a new system called AutoPPG. The system first analyzes the behaviors related to users' privacy information through various static codes and then uses natural language processing to generate interpretable sentences to describe these behaviors. To promote the generation of the Android application privacy protocol, our hope is to help users understand the behavior of the application.

Rocky Slavin et al. [30] proposed a semi-automatic framework and a tool called PVDetector to help mobile application developers check the consistency of their privacy protocol and application code and to detect the privacy protocol violation code in Android applications due to inconsistency with the description of privacy protocol. They annotate the API documents in Android SDK, construct API mapping, link the phrases in privacy protocol with API methods that generate sensitive information, and use information flow analysis to check the preprocessed privacy protocol phrases and code to find out potential violations of privacy protocol. Their ultimate goal is to provide better privacy protocol to protect the privacy of users. However, there is a problem that they rely too much on the source code of the application. Different from them, we focus on the privacy protocol document, which is easier to obtain than the source code.

Rawan Baalous et al. [29] made a semi-automatic analysis of the terms used in the privacy protocol of Android applications, automatically identified phrases from the sentences extracted from the privacy protocol, marked them with POS, found hypernym, synonym, and meronym relationships, and mapped the privacy protocol terms to the dangerous permissions of Android. The analysis results provide 128 privacy protocol terms that match Android dangerous permissions.

Wang et al. [10] studied the Android Market in China and the foreign Google market. They have shown that no one has conducted a comparative study on the application market. Therefore, they have conducted a multi-faceted and large-scale study on more than 6.2 million applications to determine the differences betweenthe Google Play Store and Android application market in China. Their research results show that China's Android application market has a greater potential crisis than Google Play Store.

According to the research of Ma et al. [31], they evaluated the privacy policy of China's mobile applications by using the content analysis method. Their research sample was 104 Chinese mobile health app privacy policy texts after three rounds of screening; Six first-level indicators (Privacy Policy attributes, personal information collection, storage, use, sharing, feedback) are constructed. On this basis, according to the requirements of "Information Security Technology—Personal Information Security Specification" (GB/T 35273-2017), the second and third level indicators are generated. Through the comprehensive evaluation of 104 apps, it is found that the score of mobile applications in user privacy protection policy is very low, and some applications have the excessive collection and abuse of user privacy data. The privacy policy of Chinese applications needs to be further improved in terms of standardization and completeness.

In addition, the research of Liu et al. [32] shows the differences in the description of application privacy protocols between China and foreign countries; Foreign Google Play Store can jump to view the privacy protocol in the app search interface for users to view; China's app market needs to download the app before it can be viewed. Therefore, in the Google Play Store, you can directly obtain the application privacy protocol documents; In the face of China's application market, after a period of development, some application markets have provided privacy protocol links on the download page. This paper crawls the privacy protocol links in Huawei AppGallery to obtain our data set.

Finally, according to the research of McDonald et al. [33], they calculated the average time spent on the monthly privacy policy in two ways and concluded that if American citizens read the content of the privacy protocol word by word, it would take an average of 40 minutes every day to read it. We believe that our successful detection of user's information-related operation sentences in the privacy protocol will greatly reduce the time spent by users reading the privacy protocol documents.

8 Conclusion and Future Work

In this paper, our research focus is based on the analysis of the Android application privacy protocol. To solve the problem that users spend a lot of time reading the privacy protocol and often ignore it, we implement the detection of sentences related to personal information operation in the privacy protocol document. We analyze the sentence composition in the privacy protocol document, use Stanford CoreNLP to segment it, and obtain its corresponding part of speech. Finally, we take the form of a dictionary as the final input. After matching the verbs, nouns, and modifiers, we finally output the sentences related to the operation of personal information. Based on the provisions on the scope of personal information, through manual verification and comparison, our precision rate, recall rate, F-score rate, and accuracy rate are 92.0%, 86.6%, 89.2%, and 92.9% respectively. We also compare our method with keyword-based retrieval. The results show that our method is better than the keyword-based retrieval method. The precision of our method increased by 44.6%. It can effectively detect the sentences related to personal information operation to help users quickly read the sentences related to personal information operation in privacy protocol documents.

In the future, we will detect whether the privacy protocol document is suspected of collecting user information beyond the scope and detect the scope of collecting and using user personal information according to different types of Android applications and the requirements of the "provisions".

Acknowledgements

This work was supported partially by the National Key Research and Development Program of China under Grant 2018YFB1403400 and the research project of Natural Science Foundation of Jiangsu Province under Grant 21KJB520029. The research project of Tongda College of Nanjing University of Posts and Telecommunications under Grant XK006XZ19013, XK004XZ19003, JG20120024.

References

[1] Y.-M.Yao, Y.-L. W, W.-Y. Jiang, Z.-Y. Wang, S. Huang, Privacy Protocol Analysis Based on Android Application, *IEEE 8th International Conference on Dependable Systems and Their Applications*, Yinchuan, Ningxia, China, pp. 632-639, September, 2021.

[2] Z.-Y. Cai, Z.-D. Wu, J.-W. Zhang, W.-Q. Wang, A BD Group Key Negotiation Protocol based on Clustering Technology, *International Journal of Performability Engineering*, Vol. 16, No. 6, pp. 875-882, June, 2020.

[3] iiMedia Research, https://data.iimedia.cn/data-classification/theme/13594980.html.

[4] China's National Bureau of Statistics, http://www.stats.gov.cn/tjsj/zxfb/202102/t20210227_1814154.html.

[5] Huawei App Market, https://appgallery.huawei.com/.

[6] Mi App Store, https://app.mi.com/.

[7] Anzhi App Store, https://www.anzhi.com

[8] Tencent App Store, https://www.android.myapp.com.

[9] Wandoujia, https://www.wandoujia.com.

[10] H.-Y. Wang, Z. Liu, J.-Y. Liang, N. V. Rodriguez, Y. Guo, L. Li, J. Tapiador, J.-C. Cao, G.-A. Xu, Beyond Google Play: A Large-Scale Comparative Study of Chinese Android App Markets, *the Internet Measurement Conference (IMC 2018)*, Boston, MA, USA, 2018, pp. 293-307.

[11] J. X. Zhang, China's National Computer Virus Emergency Response Center has detected 15 illegal mobile apps, http://edu.china.com.cn/2021-09/10/content_77744275.htm.

[12] Google Assistant, Privacy policy guide, https://developers.google.com/assistant/console/policies/privacy-policy-guide.

[13] J.-Y. Wang, M.-K. Xu, H.-Y. Wang, G.-A. Xu, Automated detection of consistence between App behavior and privacy policy of Android Apps, *Journal of Frontiers of Computer Science and Technology*, Vol. 13, No. 1, pp. 56-69, January, 2019.

[14] R. Slavin, X.-Y. Wang, M. B. Hosseini, J. Hester, R. Krishnan, J. Bhatia, T.-D. Breaux, J.-W. Niu, Toward a framework for detecting privacy policy violations in android application code, *IEEE 38th International Conference on Software Engineering*, Austin, TX, USA, 2016, pp. 25-36.

[15] P. Story, S. Zimmeck, A. Ravichander, D. Smullen, Z. Wang, J. Reidenberg, N. C. Russell, N. Sadeh, Natural Language Processing for Mobile App Privacy Compliance, *PAL: Privacy-Enhancing Artificial Intelligence and Language Technologies*, Palo Alto, CA, USA, 2019, pp. 1-9.

[16] M. M. Alani, Android Users Privacy Awareness Survey, *International Journal of Interactive Mobile Technologies* (iJIM), Vol. 11, No. 3, pp. 130-144, April, 2017.

[17] Douyin Privacy policy, https://www.douyin.com/protocols/?id=6773901168964798477.

[18] Appium, https://appium.io/.

[19] Android AAPT, https://androidaapt.com.

[20] Developers, UI Automator, https://developer.android.com/training/testing/ui-automator.

[21] J.-F. Zhou, T. Liu, L. Zou, Design of machine learning model for urban planning and management improvement, *International Journal of Performability Engineering*, Vol. 16, No. 6, pp. 958-967, June, 2020.
[22] Stanford CoreNLP, https://stanfordnlp.github.io/CoreNLP.
[23] Data.ai, How COVID-19 Has Changed Consumer Behavior on Mobile Forever, https://www.appannie.com/en/insights/market-data/covid19-consumer-behavior-mobile/.
[24] T. D. Breaux, A. Rao, Formal analysis of privacy requirements specifications for multi-tier applications, *21st IEEE International Requirements Engineering Conference (RE)*, Rio de Janeiro, Brazil, 2013, pp. 14-23.
[25] Cyberspace Administration of China, Notice on printing and distributing "the provisions on the scope of necessary personal information for common types of mobile Internet applications", http://www.cac.gov.cn/2021-03/22/c_1617990997054277.htm.
[26] Standardization Administration of the People's Republic of China, Information security technology—Personal information security specification, GB/T 35273-2020, March, 2020, http://openstd.samr.gov.cn/bzgk/gb/newGbInfo?hcno=4568F276E0F8346EB0FBA097AA0CE05E.
[27] L. Yu, T. Zhang, X.-P. Luo, L. Xue, AutoPPG: Towards Automatic Generation of Privacy Policy for Android Applications, *BT-Proceedings of the 5th Annual ACM CCS Workshop on Security and Privacy in Smartphones and Mobile Devices, SPSM 2015*, Denver, Colorado, USA, 2015, pp. 39-50.
[28] R. Baalous, R. Poet, How Dangerous Permissions are Described in Android Apps' Privacy Policies, *BT - Proceedings of the 11th International Conference on Security of Information and Networks*, Cardiff, United Kingdom, 2018, pp. 26:1-26:2.
[29] R. Slavin, X. Wang, M. B. Hosseini, J. Hester, R. Krishnan, J. Bhatia, T. D. Breaux, J. Niu, PVDetector: a detector of privacy-policy violations for Android apps, *International Conference on Mobile Software Engineering and Systems* (MOBILESoft '16), Austin, Texas, USA, 2016, pp. 299-300.
[30] A. Oglaza, R. Laborde, P. Zaraté, Demonstration of KAPUER: A privacy policy manager on Android, *BT-13th IEEE Annual Consumer Communications & Networking Conference, (CCNC 2016),* Las Vegas, NV, USA, 2016, pp. 282-283.
[31] C.-Y. Ma, Q.-K. Liu, Research on the Privacy Policy's Evaluation and Empirical Study of Mobile Health Applications, *Library and information service*, Vol. 64, No. 7, pp. 46-55, April, 2020.
[32] J. Liu, J. Bai, A Comparative Study on the Privacy Policy of Chinese and Foreign Mobile APP, *Journal of Shantou University (Humanities & Social Sciences Edition)*, Vol. 33, No. 3, pp. 82-87, March, 2017.
[33] A. M. McDonald, L. F. Cranor, The cost of reading privacy policies, *I/S: A Journal of Law and Policy for the Information Society*, Vol. 4, No. 3, pp. 543-568, January, 2008.

Biographies

Yongming Yao received the B.S. degree in communication engineering from Nanjing University of Posts and Telecommunications in 2010 and the M.S. degree in computer system architecture from Xi'an University of Posts and Telecommunications in 2013. He is currently pursuing the Ph.D. degree in software engineering at Army Engineering University of PLA. He has been an Assistant professor with the software engineering department, Tongda College, Nanjing University of Posts and Telecommunications. His research interests are in the area of crowdsourced software testing and Android permissions detection.

Yulin Wang received the B.S.degree from the Tongda College, Nanjing University of Posts and Telecommunications in 2017. He is currently pursuing the M.S.degree in Shandong Institute of business and technology.

Weiyi Jiang received the B.S.degree from the Tongda College, Nanjing University of Posts and Telecommunications in 2017. He is currently pursuing the M.s.degree in Nanjing University of Information Science and Technology.

Ziyuan Wang received the B.S. degree in mathematics and the Ph.D. degree in computer science from Southeast University, Nanjing, China,in 2004 and 2009, respectively. He is currentlyan associate professor in computer science with the School of Computer Science and Technology, Nanjing University of Posts and Telecommuni-cations, Nanjing, China. He was a postdoctoralresearcher with the Department of Computer Sci-ence and Technology, Nanjing University, Nan-jing, China. His research interest mainly focuses on automated softwaretesting techniques.

Song Huang received his Ph.D. degree from PLA University of Science and Technology. He is a member of CCF and ACM. He is currently a professor of software engineering at Software Testing and Evaluation Center from Army Engineering University of PLA. He is a member of the advisory boards of Journal of Systems and Software, IEEE Transactions on Reliability, etc. He has contributed more than 100 journal articles to professional journals. His current research interests are in the areas of software testing, quality assurance, data mining, and empirical software engineering.