A Password Strength Evaluation Algorithm based on Sensitive Personal Information

Xinchun Cui[1,2]
1.School of Computer Science
Qufu Normal University
Rizhao,China, 276826
cxcsd@126.com

Xueqing Li
1.School of Computer Science
Qufu Normal University
Rizhao,China, 276826
lixueqing_98@163.com

Yiming Qin
1.School of Computer Science
Qufu Normal University
Rizhao,China, 276826
qym981007@outlook.com

Ding Yong
2.Guangxi Key Laboratory of Cryptography and Information Security
Guilin, China, 541004
stone_dingy@126.com

***Abstract*—Many Internet service providers are still using traditional password strength evaluation methods, resulting in user passwords being vulnerable to social engineering attacks. We believe that the password strength evaluation method based on sensitive personal information has great research value for improving the security of password authentication system. In this paper, we use the structure segmentation algorithm and the bidirectional matching algorithm to investigate how users' personal information is used in passwords. Then, we present a sensitivity personal information coverage evaluation function that represents the correlation between users' password and their personal information. Finally, a password strength evaluation method based on sensitive personal information is proposed. This method is composed of three stages: preprocessing stage, prediction dictionary generation stage and password strength evaluation stage.**

Keywords—information security, password authentication, password strength evaluation method, data mining

I. INTRODUCTION

In recent years, information technology represented by mobile Internet and e-commerce has greatly facilitated people's lives, and information security issues closely related to the Internet have received more and more attention. As an important way to protect the security of user information, password authentication is widely used in major Internet service providers.

In the password authentication system, the system requires the user to create a string as a means of verifying the identity of the user. However, a study by J. Bonneau[1] found that users tend to set simple passwords for convenience, such as grouping names, phone numbers, or birthdays to form an account password. In order to improve the security strength of the user's password, the general Internet service provider will enforce the Password Strength Metric (PSM) to feedback the strength of the password to the user, and propose a high-strength password to the user according to the evaluation result[2-3]. However, at present, various Internet service providers do not give password strength feedback results under the premise of researching user personal information, resulting in a large security risk of user passwords, posing a potential threat to user information. In order to better protect the security of user passwords, researchers at home and abroad have done a lot of research on password strength evaluation methods in recent years, and gradually developed the rule-based password strength evaluation method, the password evaluation method based on pattern detection and the password strength evaluation method based on the attack algorithm[4].

The rule-based password strength evaluation method simply evaluates the password strength by detecting the password length and the type of characters contained in the password[5]. Many websites use this method for password evaluation, such as the password evaluation method used by NetEase mailbox when registering: (1) If the password length is greater than or equal to 6, but all characters are similar characters, the password strength evaluation level is weak; (2) If The password length is `greater than or equal to 6, and the password strength evaluation level is medium if there are two types of characters; (3) if the password length is greater than or equal to 6, and contains three or more characters, the password strength evaluation level is strong, as shown in Table I. Although the rule-based password strength evaluation method is simple and feasible, it can be seen from the above test that with the evaluation method, the password strength that satisfies the rule but is very regular and easily cracked cannot be correctly fed back, and the evaluation result has low reliability.

TABLE I. NETEASE MAILBOX REGISTRATION PASSWORD AND STRENGTH EVALUATION LEVEL

Password	Strength Evaluation Level
123456	Weak
abcdefgh	Weak
12345@	Medium
12345a@	Strong

The password matching method based on feature matching refers to evaluating the password strength by detecting whether the password contains some fixed string. The pattern matching fonts mainly set by this method include common semantic matching, sequential arrangement characters, keyboard layout and weak password dictionary set. The password strength is finally obtained by performing a weighted summation calculation on each of the matched strings. Although this method is more rigorous than the rule-based password strength evaluation method, its evaluation results are very dependent on the weight designer and have great subjectivity[6].

The password strength evaluation method based on the attack algorithm refers to attacking the password by using an attack algorithm, and the evaluation value is given according to the anti-hacking ability of the password[7]. Although the results obtained by this evaluation method are more realistic and accurate, different attack algorithms may have different password strength values for the same password, and it is difficult to define the true strength of the password. In addition, in real life, the calculation time of this method generally

978-1-6654-0392-4/20/$31.00 ©2020 IEEE
DOI 10.1109/TrustCom50675.2020.00211

exceeds the maximum tolerance time of user registration, and has no universality.

Research shows that the evaluation method that can correctly feedback the password strength can effectively improve the security of the password[8]. The irrationality of the evaluation results can be misleading to the user, causing the user to set a weak strength password, resulting in leakage of privacy data. Therefore, in the current network environment, the unreasonable password strength evaluation strategy is an important cause of password vulnerability. The ultimate goal of this paper is to propose a password strength assessment method that accurately reflects the strength of the password. The innovation of this method is that it evaluates the password strength based on sensitive personal information.

In this paper, we first conduct statistical analysis on large-scale password leakage data, and study the basic characteristics of the password character composition distribution of the data set. Based on integrated password data and personal information data, we studied the vulnerability configuration behaviors such as the use of personal information in passwords. Then, a new metric is proposed, the sensitive personal identity information coverage ratio α, used to represent the relationship between personal information and user password. We define the sensitive personal identity information coverage ratio α to quantify the correlation between user password and personal information. Degree, and a password strength evaluation function is proposed. Finally, a password strength evaluation method based on user-sensitive personal identity information is proposed. The method is divided into three stages: pre-processing stage, prediction dictionary generation stage and password strength evaluation stage.

II. Sensitive Personal Information in User Passwords

Traditional research has shown that users often use simple dictionary words to construct passwords, such as keyboard layout strings such as "qwerty" and "qweasd" or semantic strings conforming to the user's first language, such as "iloveyou" and "5211314" [9-10]. However, people now tend to create a password based on their personal information, and traditional password strength evaluation methods can no longer provide high-confidence evaluations. Understanding the use of passwords and personal information can help us further improve the security of our passwords.

A. 12306 Data Set

In recent years, some password data sets have been exposed to the public, and there are many password analysis studies based on these data sets[11]. In this article, a leaked password data set from the website www.12306.cn will be used to illustrate the relationship between sensitive personally identifiable information and passwords.

1) 12306 Data Set: The data set was leaked to the public by anonymous attackers at the end of 2014 and contained approximately 130,000 passwords. In addition, several types of sensitive personally identifiable information such as name, identification number (ID number, passport or other number), telephone number and email address are included. This information can be used to distinguish or track an individual's identity. Since the real user registration ID number exists in the data set, the information of the data set is determined to be reliable[12].

2) Basic Analysis of Passwords: In order to maintain data consistency, the data set needs to be cleaned and processed. We first convert all Chinese names into Chinese Pinyin, and then directly delete the user data whose ID number is not equal to 18 digits. Data cleaning process always contains 131,388 passwords. The analysis shows that the average length of the data set password is 8.438, and the top 10 passwords in the data set are shown in Table II. It mainly includes weak passwords (such as 123456, 111111, etc.), keyboard layout passwords (such as 1qaz2wsx, etc.) and Chinese Pinyin passwords. Then we use the LDS method (L means Letter, D means Digit, S means Special character) [13] to divide the password, and the distribution of characters in the password is counted as shown in Table III. According to the data, the Chinese prefer to use English letters and numbers as the main components of the password, and rarely use special characters.

TABLE II. Top 10 Passwords in The Data Set

Rank	Password	Amount	Percentage
1	123456	389	0.296%
2	a123456	280	0.213%
3	123456a	165	0.126%
4	5201314	160	0.122%
5	111111	156	0.119%
6	woaini1314	134	0.102%
7	qq123456	98	0.075%
8	123123	97	0.074%
9	000000	96	0.073%
10	1qaz2wsx	92	0.070%

TABLE III. Character Distribution in Passwords

Structure	Amount	Percentage
Letter	363433	67.203%
Digit	745073	32.780%
Special character	191	0.017%

B. Analysis of Password Structure Based on Personal Information

In order to better analyze the relationship between personal information and password, we improved the LDS method of the password segmentation training phase in the Probabilistic Context-Free Grammar (PCFG) [14], adding five sensitive personal information flag segments. as shown in Table IV. These five fields exist independently as the composition of the password, and the matching priority is higher than L, D and S. For example, the password "19870102zhang" is divided into $[B]_8[N]_5$.

TABLE IV. Sensitive Personal Information Segments

Information Type	Description
Birthday[B]	birthday
Name[N]	Name (including full spell, initials, etc.)
Telephone[T]	phone
Email[E]	email address
Identity-card[ID]	Authentication number

The name matches the pinyin of the user's name, ignoring case. For example, if Zhang San is registering, his password is “19870206zhang”. The "zhang" in the password is included in the pinyin "zhangsan" and the password is considered to contain the name. The birthday match is to divide the user's birthday information field into YYYY, MM, M (single month, 02 can also be extracted as 2), DD, D (single date, 06 can also be extracted as 6). Among them, YYYY matches the year, MM or M matches the month, and DD or D matches the date.

If the password contains birthday information such as 19870206, 0206, etc., these characters will be recognized as the birthday type. Telephone type matching was extracted using a 3-4-4 distribution. If the mobile phone number is 12345678910, then 123, 4567, 8910 is the telephone type. Mailbox matching is to divide the mailbox string into two parts by @, one part is the mailbox user name, and the other part is the mailbox service provider. The shape is like 12344321@qq.com will be divided to get a list of two fields 12344321 and qq. The ID card number field takes the last four digits as the identity. The structure segmentation algorithm based on personal information is as shown in Algorithm 1.

Algorithm 1 Structure Segmentation Algorithm Based on Personal Information

```
Procedure GetInfoList(SPInfo)
    List←empty_ List<String>
      [N],[B],[ID],[E],[T]←empty_String[]
      getNameList(Name):
        String←empty_String[]
        String←Name.split("\\s+")
        String←First letter of the Name
        for EachNameString∈String
            InfoList←add[N]
      getBirthdayList(Identity-card):
        birthday = Identity-card.substring(6,14)
          [B]=birthday.substring(0,4)&birthday.substring(4,6)&birth
          day.substring(6,8)
          InfoList←add[B]
      getIdList(Identity-card):
        [ID]= Identity-card.substring(14, Identity-card.length())
        InfoList←add[ID]
      getEmailList(Email):
        [E]=email.substring(0 to '@')&email.substring('@' to '.')
        InfoList←add[E]
      getTelList(Telphone):
        [T]= Telphone.substring(0,3)
        &Telphone.substring(3,7)&Telphone.substring(7,11)
        InfoList←add[T]
  return InfoList
end procedure
```

Next, we store the personal information field after the segmentation algorithm in the list and wait for it to match the password. Based on the above analysis method, the bidirectional matching algorithm for password and personal information is proposed, as shown in Algorithm 2. According to the bidirectional matching algorithm , the dataset counted 90,619 passwords containing personal information. The occurrence probability of each information type is shown in Table V. Through probabilistic analysis, the type of birthday and name appears most frequently in the password, while other types appear less frequently. Therefore, the data can be used as the weight of each information type in the password as shown in the last column of Table V.

Algorithm 2 Bidirectional Matching Algorithm

```
Procedure Match(Password,InfoList)
Substring←get_all_Substring(Password)
  for EachString∈Substring do
    if Match_B(EachString,InfoList) then
      tag←"[B]"
      getlength_B=this.EachString
      remainder←Password.split(EachString)
      break
    end if
    if Match_N(EachString,InfoList) then
      tag←"[N]"
      getlength_N=this.EachString
      remainder←Password.split(EachString)
      break
    end if
    ...
    if Match_ID(EachString,InfoList) then
      tag←"[ID]"
      getlength_ID=this.EachString
      remainder←Password.split(EachString)
      break
    end if
  end for
    if remainder.size()>=1 then
      for i<remainder.size() do
        if eachremainder∈complete_B || eachremainder∈
complete_N ||
          eachremainder∈complete_T || eachremainder∈
complete_E ||
          eachremainder∈complete_ID do
          tag←Corresponding tag
          getlength_ Corresponding tag=this.eachremainder
          break
        end if
      end for
  end if
  results←extract_structures(new form)
  return results
end procedure
```

TABLE V. PASSWORD STRUCTURE DISTRIBUTION BASED ON PERSONAL INFORMATION

Information Type	Amount	Ratio of Occurrence	Weights
Birthday[B]	71445	0.788	0.788
Name[N]	36020	0.397	0.397
Email[E]	7207	0.079	$\frac{Length\ (flag\ segment)}{Length\ (password)}$
Telephone[T]	1140	0.013	
Identity card[ID]	686	0.008	

III. CORRELATION QUANTIFICATION

A. Abbreviations and Acronyms

Sensitive personal information coverage refers to the proportion of the personal information field in the password, which typically ranging from 0 to 1. A coverage ratio of "0" indicates that no personal information is included in the password, and the closer the coverage ratio is "1", the more personal information the entire password contains, and the personal information has a stronger correlation with the password. L means the password length. N means the number of matching information type. The length of each personal information type in the password is P. The weight K is assigned to different types of personal information appearing according to the research data. The calculation formula of coverage ratio α is shown in formula (1).

$$\alpha = \sum_{i=1}^{N} \frac{P_i}{L} K_i \tag{1}$$

To illustrate how the coverage is calculated, we assume a user named Zhang San, whose birthday is February 6, 1987, and his password is "19870206zhang". Then, according to the bidirectional matching algorithm, his password structure is $[B]_8[N]_5$. Obviously, this password is highly correlated with personal information, then the sensitivity of this password is calculated as follows:

$$\alpha = \sum_{i=1}^{2} \frac{P_i}{L} K_i = \frac{8}{13} \times 0.788 + \frac{5}{13} \times 0.397 \approx 0.638 \tag{2}$$

B. Password Strength Evaluation Function

The strength evaluation function is shown in formula (3), the strength value of the above password can be obtained as

0.362. According to the conclusion of many experiments, it is concluded that the password strength is weak when the password strength value is lower than 0.6.

$$f(x) = 1 - \alpha \tag{3}$$

IV. Password Strength Evaluation Method Based on Sensitive Personal Information

This paper starts with sensitive personally identifiable information and conducts research using a known password data set containing personal information. Under the premise that the weak password in the general dictionary has been detected, the password strength evaluation method based on sensitive personal information is summarized. This method includes three stages of preprocessing, prediction dictionary generation and password strength evaluation, as shown in Fig 1.

The task of the preprocessing stage is to lay a theoretical foundation for the subsequent work based on the password structure of a large number of password data sets.

The main work of the prediction dictionary generation stage is to use the construction conclusion obtained by the preprocessing to divide the sensitive personal information input by the user, and obtain a prediction password dictionary containing the personal information field. It is convenient to match the password to be detected input by the subsequent user, and calculate the strength value of the password.

In the password strength evaluation stage, the previously obtained research data is used to propose a password strength evaluation method based on sensitive personal information coverage. The strength value of the user's current password is calculated by the degree of matching of the user password with the predicted password dictionary.

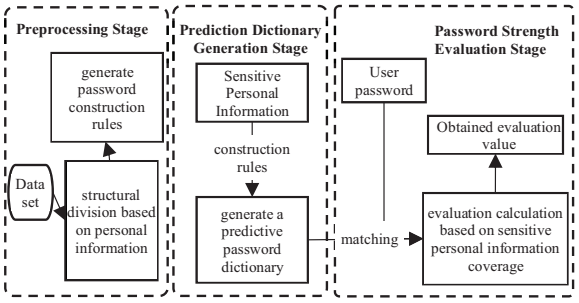

Fig. 1. Password Strength Evaluation Method Based on Sensitive Personal Information

V. Summary and Outlook

This paper deeply analyzes the user password structure based on sensitive personal identity information, summarizes the password composition rules. At the same time, it defines the coverage ratio of the personal information in the calculation password, and obtains the quantified password security strength value through the evaluation function. Finally, we propose a password strength evaluation method based on sensitive personal information coverage. The advantage of the model proposed in this paper is that the data analysis method is used to find the intrinsic relationship between password and personal information from the mass password data set, and the accuracy and authenticity of the strength evaluation result are improved.

The next step is to update the weights of various types of personal information in this method by obtaining more data sets to summarize more precise rule data. It is also possible to study the password suggestion change strategy after applying this method, so that users can avoid low intensity. The method of creating passwords to enhance password security.

Acknowledgment

This work is partially supported by National Nature Science Foundation of China (71971190), High-quality Course for Graduate Education of Shandong Province (Digital Image Processing, SDYKC19178), Shandong Social Science Planning Research Project (20CSDJ20)，and Guangxi Key Laboratory of Cryptography and Information Security (No.GCIS201903).

References

[1] Bonneau J. The Science of Guessing: Analyzing an Anonymized Corpus of 70 Million Passwords[C]. Security & Privacy. IEEE, 2012.

[2] Kong Xiangqian, Shao Wenwu. Methods and Application of Password Strength Met[J]. China Computer & Communication, 2019,31(21):51-52+55.

[3] Chen Ying, Hong Di, Yanan Liu. Password Strength Evaluation Method Based on Probabilistic Context-Free Grammar[J]. Internet of Things Technologies, 2017,7(04):59-61.

[4] Ur B, Kelly P G, Komanduri S, et al. How does your password measure up? The effect of strength meters on password creation[C]// Proc of SEC 2012. Berkeley, CA: USENIX Association, 2012: 65-80.

[5] Weir M, Aggarwal S, Collins M, et al. Testing metrics for password creation policies by attacking large sets of revealed passwords[C]// Proc of ACM CCS 2010. New York: ACM, 2010:162-175.

[6] Yan Ruirong, Password Strength Meter Method and Software[D]. South China University of Technology, 2018.

[7] Wang Ping, Wang Ding, Huang Xinyi. Advances in Password Security[J]. Journal of Computer Research and Development, 2016,53(10):2173-2188.1

[8] Egelman S, Sotirakopoulos A, Muslukhov I, et al. Does my password go up to eleven?: The impact of password meters on password selection[J]. 2013.

[9] M. E. Hellman, "A cryptanalytic time-memory trade-off," IEEE Trans. Inf. Theory, vol. 26, no. 4, pp. 401–406, Jul. 1980.

[10] Li Y, Wang H, Sun K. Personal Information in Passwords and Its Security Implications[J]. IEEE Transactions on Information Forensics and Security, 2017:1-1.

[11] Kelley P, Komanduri S, Mazurek M L, et al. Guess again (and again and again): Measuring password strength by simulating password-cracking algorithms (CMU-CyLab-11-008)[J]. 2011.

[12] Li Y, Wang H, Sun K. A study of personal information in human-chosen passwords and its security implications[C]// Proc of the 35th Annual IEEE International Conference on Computer Communications. Piscataway, NJ: IEEE Press, 2016: 1-9

[13] Zhang Meng-li, Zhang Qi-hui, Liu Wen-Fen, Hu Xue-xian, Wei Jiang-Hong. A Method of Password Attack Based on Structure Partition and String Reorganization[J]. CHINESE JOURNAL OF COMPUTERS, 2018:1-16

[14] Ma J , Yang W , Luo M , et al. A Study of Probabilistic Password Models[C]// 2014 IEEE Symposium on Security and Privacy. IEEE, 2014.