Non-memorizing evolutionary authentication algorithm using the preference symbols for secondary authentication

KwangCheol Rim*
Chosun University
309, Pilmun-daero, Dong-gu, Gwangju, Korea
rim1201@hanmail.net

Abstract

In today's diverse society, the need to authenticate system connections and management authority is continuously rising. Using the same password in all authentication processes is dangerous. Numerous methods have been proposed for document security, document management, and minimization of risk due to insiders. All these methods require secondary authentication for important documents. Since users can only create passwords within the range of human cognitive intelligence, these are vulnerable to attacks. This paper proposes an algorithm that generates passwords without requiring memorization. It examines existing non-memorization type algorithms, and proposes an algorithm that solves the disadvantages, of the inconvenience of new hardware and the cost of implementing new infrastructure, afflicting existing algorithms. Initial data regarding individual preferences were collected, and a statistical authentic method using these data was designed. We first examined the coincidence of preference symbols through a survey in order to ensure statistical safety. The preference symbols authentication algorithm was proposed based on the results of the survey. Our proposed system for secondary authentication in internal security without memorization is convenient for users, and provides excellent statistical security against known attacks.

Keywords: Preference Symbol, Secondary Authentication, Non-memorizing Password, Internal Security

1 Introduction

Given the proliferation of Information and Communication Technologies (ICT) in recent times, people are able to conveniently access information over the limits of time and space. This requires that users understand and manage a large number of authentication factors. An operation requiring multiple user access may cause illegal accesses. Simply using different passwords for different services is not a satisfactory solution to the problem.

The average number of passwords that can be remembered within the range of a user's common cognition capability is approximately three. In reality, however, the average user is registered to more than 10 sites and has to remember approximately 20 passwords, including those for financial corporations and public offices. Hence, most users end up using the same ID and password for various online accounts and services. Generating a new ID and password each time and recording it in a notebook or a memo pad may be convenient for self-authentication but poses difficulties with regard to managing passwords. Another method proposed using only one access terminal, but this solution is rendered obsolete in the current, diversified smart communication environment [4].

A cyber-attack called a "list attack," which has evolved from a "dictionary attack", is widely used as the typical pattern for illegal access. In a list attack, an attacker obtains a list of IDs and passwords leaked from a certain service, and then tries to access other sites using these. It is well known that passwords

Research Briefs on Informaiton & Communication Technology Evolution (ReBICTE), Vol. 1, Article No. 13 (January 15, 2015)

[*] Corresponding author: Room 501, 5F, 1385-6 beonji, youngbongdong, bukgu, Kwangju, Korea, Tel: +82-625302030

can be cracked by using a computer equipped with general-purpose computing on graphics processing units (GPGPUs), even if they are encrypted using a hash function such as MD5 or SHA-1. To prevent illegal access, authentication-enhancing methods involving multiple verification measures are used in the industry [2].

On May 22, 2013, Twitter announced that it was providing two-factor authentication using a one-time password as an optional function to prevent illegal access. Google and Dropbox have also introduced this service, whereas Evernote is in the process of implementing it. In 1999, the Japanese mobile phone operator NTT DoCoMo implemented an authentication method that connects a terminal and a network using terminal authentication instead of ID and password. However, as smart communication has advanced, open operating systems (OS), multi-device technologies, mobile networks, WiFi, etc. have emerged. Thus, the value of it dropped rapidly, causing the industry to return to authentication using an ID and password.

The U.S. Department of Defense uses authentication that sets a "Strong Key" by combining special characters, alphabets, and numbers. However, because it only emphasized security and was not in accord with user preferences and characteristics, users began to share their passwords with others for the sake of convenience.

In November 2011, Richard Guidorizzi suggested five modified forms of "Jane123," a hand-written password form, at a symposium of the U.S. Department of Defense. However, it was found that this form could be easily cracked. Following this, he proposed a method that authenticates users based only on their behavioral characteristics by using a user movement, relative to the direction in which he/she is seated in front of a terminal as a "cognitive footprint," and has been developing it. However, such algorithms have the disadvantage of requiring a terminal that can sense the user's movement and handwriting; this requires large amounts of movement data and handwriting recognition data as a precondition for authentication [3].

Roy Maxion at Carnegie Mellon University studied "keystroke dynamics," which considers the time for which a user "dwells" on a specific key as well as his/her movement time from a specific key to another. Maxion proposed an authentication method based on keystrokes. Maxion suggested that users dwell on a key or move to another in 100 ms on average, and argued that on average 10 ms error occurred if a specific attacker carries out keystroke by imitating the user due to slowing down or quickening keystroke. The commercialization of this method is expected to take a considerable amount of time, considering that it requires a large number of prerequisite keystrokes and that a blink of an eye is only 275 ms long, which is a short interval of time [9], [5].

Multiple access control is applied to document security and internal networks as a basic security measure. Authentication processes occur several times during this, such as system connection authentication and management right authentication. Internal document security can be divided into Digital Right Management (DRM), server-based computing (SBC), and data loss prevention (DLP) [7], [6], [8].

The concept of internal security focuses on document security against outsider attacks, insider betrayal, and the hacking of an insider's terminal. Multiple passwords are required to complete hierarchical authentication, such as for system connection and authorizing access to documents. In this regard, memorizing each and every password used for internal security causes a major inconvenience for the user.

This paper proposes a non-memorizing password for user authentication. Authentication using biometric information or keystrokes requires new hardware, whereas authentication using a preference symbol is a password-based system that employs the user's intelligence and does not require additional equipment. We design a usable password system at the second authentication step for internal document security. In Section 2, we examine statistical surveys on satisfying authentication requirements. In Section 3, the evolutionary preference symbol authentication algorithm based on the results of statistics is presented. We test our system's performance in Section 4, and conclude this paper in Section 5.

Table 1: Comparison of the internal security methods

	Advantages	Disadvantages
DRM	· Limits use authority by encrypting documents · Protects reading of keystrokes and makes it possible to trace documents even if they are leaked · Limits the circulation period and the number of users allowed to read each document	· Does not protect against data leaks by providing encryption. · Possibility of loopholes in security during the period of new connection establishment.
SBC	· Environment that leave 100% of all users' OS, applications and information on a server and use users' PC for only show the execution results · Can be used to prevent leakage at the source because all data do not exist at the local system.	· Decrease in execution speed · Big initial investment required.
DLP	· Contents cognition type security detecting solution provides the monitoring, protecting and reporting functions for classified data · Protects leakage of secret files while allowing media usage and immediately reporting breaches of policy	· The structure that the data are stored is on the user's local disk · The risk of data leakage still exists · System load increases due to the continuous searching and monitoring

2 Preference symbols

We first examined the range of cognitive intelligence required for the preference symbols. To gauge the stability of the preference symbols, 150 students of Chosun University were surveyed regarding their preferences.

For a month, we repeatedly inquired about the subjects' preferences. The ratio of men to women among subjects was maintained at 5 to 5, and is representative of the general population. The survey questionnaire was as follows:

The survey targeted students in the departments of civil engineering and environment engineering at Chosun University. A total of 118 students, excluding insincere respondents, replied to the survey. The first survey was conducted on April 7, 2014 and the second on May 5, 2014. To prevent the error of memorizing questionnaires, the second survey was justified by informing the subjects that responses to the first survey had been lost.

The coincidence rate of each question for all respondents was 74.64%. As is evident from data listed in Table 3, responses to the question concerning the subjects' favorite school showed a high association with their personal information, with a coincidence rate of close to 100%. The lowest coincidence rate was 58.47% for the question regarding the subjects' favorite person. The average coincidence rate for the six lowest-rated responses was 62.29%. According to this, if the coincidence rate of six randomly selected questions was checked, responses to at least four questions would coincide.

Table 2: Survey questionnaire for preference symbols

Student No.		Name	
Please write down your preference for each item.			
Elementary school		Nation	
Middle school		Bird	
High school		Person	
Color		Place	
Food		Sports	
Mountain		Game	
Pet		Company	
Flower		City	
Number		Weather	
River		Subject	

Table 3: Coincidence rate according to the preference symbol questionnaire

	Coincidence	Ratio (%)		Coincidence	Ratio (%)
Elementary school	114	96.61	Nation	86	72.88
Middle school	116	98.30	Bird	70	59.32
High school	116	98.30	Person	69	58.47
Color	96	81.35	Place	74	62.71
Food	78	66.10	Sports	82	69.49
Mountain	83	70.33	Game	86	72.88
Pet	101	85.59	Company	89	75.42
Flower	75	63.55	City	79	66.94
Number	92	77.96	Weather	97	82.20
River	75	63.59	Subject	86	72.88

3 Evolutionary authentication algorithm using preference symbols

As mentioned in Section 2, our subjects' preferences coincided to up to 80%. This shows that people can recognize personal information, such as their favorite things and places, without memorizing them. Personal information is naturally committed to memory; people do not have to try to remember it.

The greatest disadvantage of password-based authentication systems is that they are unstable against exposure. A single instance of data exposure can destroy the entire system. Frequently changing passwords by periodically generating a new one is a solution, but it incurs the inconvenience of memorization. Since the user of the password system is human, he/she is not free of the range of cognitive capability and memorization ability. For the sake of password protection, methods for creating strong passwords advise the inclusion of special characters, numbers, and capital letters in a password. This technique helps prevent dictionary or list attacks against a system, but is still vulnerable to exposure. Current password systems are still vulnerable attackers who infiltrate an insider's terminal using hacking tools or cracks. In this paper, we design a preference symbol-based password authentication system using personal information, which is a kind of cognitive fingerprint. The system is composed of information storage, question generation, authentication, and a user terminal. Personal preference symbols are stored during information storage using initial questionnaires to generate a password. The information to be used for passwords is gradually updated by having about two "incorrect answer correction" questions be stored as new information using the user's eye blink time. This is a variable password generation system, and creates a new password whenever a user connects to the system.

To improve safety, a question with a comparatively high coincidence rate, such as the subject's alma mater or favorite animal, is included at the question-selection stage. Incorrect answers are saved during information storage, and the newly entered preference symbol is derived when the next question is selected, and this leads to an increase in the coincidence rate at a certain point. In other words, our proposed method is designed as an evolutionary password system where the coincidence rate gradually increases whenever a connection is made. The matching rate, at around 70%–80% at the initial connection, increases gradually while authentication continuously progresses.

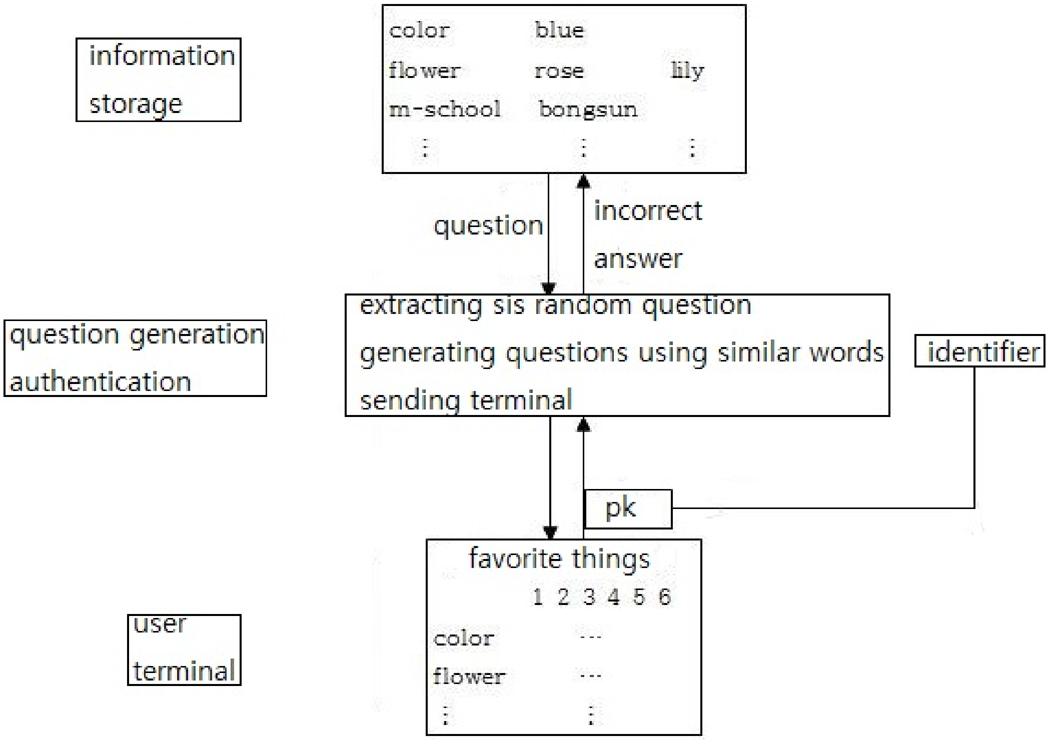


Figure 1: Structure of non-memorial password system

The basic components of the system are as follows:

- Information storage: It stores initial personal preference information and connection change information.
- Question generation and authentication: It generates similar questions by using personal preference symbols randomly extracted from information storage. Following that, it displays them on a user's terminal and completes personal authentication by using the responses selected by the user.
- User terminal: It displays questions transmitted by the question generator on the screen, and ensures keyboard security by making the user touch or clicks the input buttons.

3.1 Information storage

The initially registered preference symbols are encrypted and stored in the information storage. Following this, the new preference symbols generated from a terminal are stored at the second part of the memory space in a record. The preference symbols generated at the third instance of connection are stored by discarding the first set and moving the dataset in the second data section to the first.

As shown in Table 4, recorder A describes the kind of preference symbol. The initially generated data is entered into recorder B. An incorrect answer selected from the generated data at a terminal by the question generator is entered into recorder C. All data are encrypted and stored. Recorder C of the relevant number is decrypted and sent if a request from the question generator is received. The data of recorder B is decrypted and sent when recorder C is empty.

Table 4: Data in the information storage

	A	B	C
1	flower	rose	lily
2	color	blue	-
3	m-school	bongsun	-
⋮	⋮	⋮	⋮

Incorrect answers generated by the user are stored when questions are generated at the question generator and authentication is completed through the identifier. The incorrect answers are immediately stored at recorder C when it is empty. If recorder C contains data, these are shifted and stored in the order C → B → A. The data in recorder A is discarded.

3.2 Question generation and authentication

The question generator randomly extracts six values from the data stored in the information storage. It generates the questions by using generalized similar words according to each extracted category. It stores the answers to the six categories and sends the questionnaires to a user's terminal.

The answers selected by a user at a user terminal are encrypted with the public key and sent to the identifier. If more than four answers coincide with those in the identifier, the user is admitted as a legitimate user, and the authority for use is given. Data regarding incorrect answers in the question generator are sent to the information storage and stored as a new data. All tasks, excluding those involving a user terminal, are carried out on our propsed system that contains the information storage. The order of granting authority is as follows:

① Extracting six random data values from information storage

② Generating questions using similar words

③ Storing answers to questions in the identifier

④ Sending the generated questions to a user terminal

⑤ Selecting a terminal and comparing responses from it with those in the identifier after decrypting answers

⑥ Granting access and storing data regarding the incorrect answers into information storage

We designed our system so that two data values can simultaneously be used in information storage as an example when the question generator creates questions. It has the consequence that data in information storage are integrated into one of the data sets depending on whether type A or B of the data in the information storage is selected.

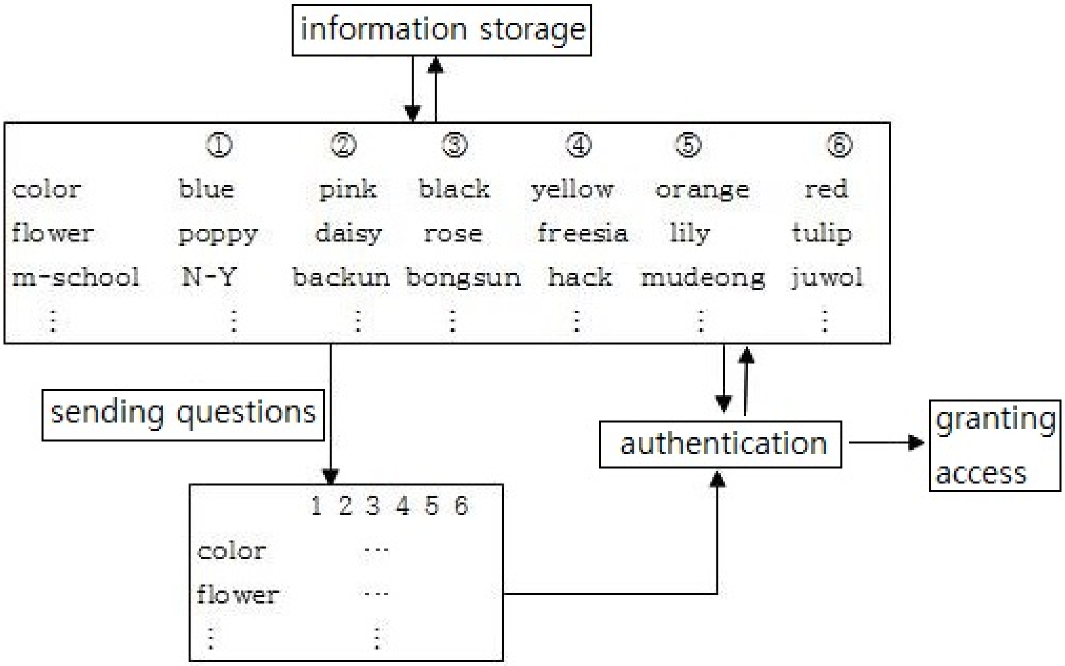
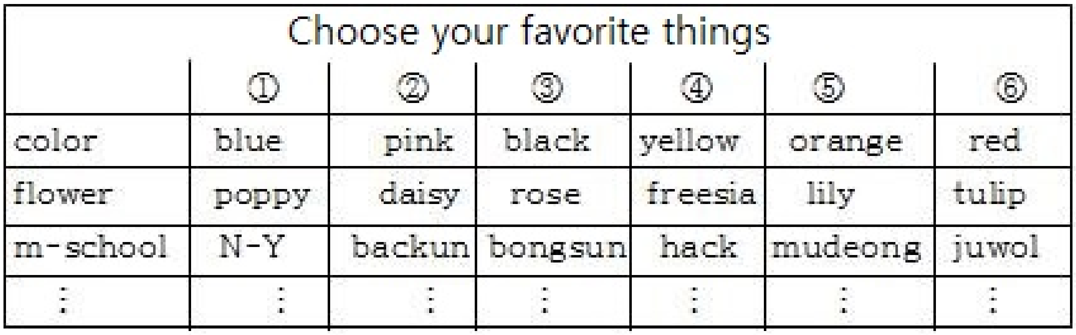

information storage
① ② ③ ④ ⑤ ⑥
color blue pink black yellow orange red
flower poppy daisy rose freesia lily tulip
m-school N-Y backun bongsun hack mudeong juwol
⋮ ⋮ ⋮ ⋮ ⋮ ⋮ ⋮
sending questions
authentication
granting access
1 2 3 4 5 6
color ···
flower ···
⋮ ⋮

Figure 2: Question generation and authentication

3.3 User terminal

The questions transmitted from the question generator are displayed on a user screen. The screen is assumed to satisfy the basic condition that it is not visible to others. A mouse or a touch pad is used to select numbers to reduce the risk of hacking or cracks. The six selected questions are sent if a user's selects his/her pre-recorded preference out of the examples of questions shown on the screen and the number or the example sentence of the question is selected. The selected numbers are sent to the identifier of the question generator after encrypting it using public key cryptography techniques. If more than four of the sent answers coincide with the stored answers in the identifier, the user is admitted as a legitimated user, and the authority is given.



Choose your favorite things

	①	②	③	④	⑤	⑥
color	blue	pink	black	yellow	orange	red
flower	poppy	daisy	rose	freesia	lily	tulip
m-school	N-Y	backun	bongsun	hack	mudeong	juwol
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Figure 3: Terminal screen

Let us examine the security of a user terminal under the assumption that the system is internally secure. If a user terminal is attacked by unscrupulous sources, the numbers for the questions selected by the user currently in effect are exposed. However, the same number cannot be used for authentication at the next connection attempt. Thus, even if a user terminal is compromised by a crack rooted by an attacker, an attack is not possible even if he/she knows the user's preference symbol questions and the preference symbols. Thus, our proposed system offers greater safety than keyboard input by using a touchscreen or a mouse.

4 Verification

Biometric information, keystrokes, mouse pattern recognition, etc., are being studied as non-memorizing password systems. Biometric information authentication is an authentication method that uses fingerprint or iris recognition, whereas mouse pattern recognition involves discovering a user's unconscious mouse patterns by letting him/her play a game for approximately 30 minutes. In this paper, the mouse pattern and biometric information authentication are excluded from comparison with our proposed system because of their implementation costs and authentication timing.

Keystroke authentication is a method that involves storing an appropriate quantity of keyboard inputs at the initial stage and carrying out authentication by comparing the saved keystroke information with a user's information when he/she requests authentication.

Table 5: Comparison of the keystroke and the preference symbols authentication

	Initial number of the input characters	Required number of characters for authentication	Match rate
Keystroke	4000	650	60%
Preference Symbols	90	6 (number)	74%

Buch et al. [1] entered a 4,000-character string as the initial value and a fixed 652-character string, or a character string of any length greater than 650, at the time of authentication. Following this, they designed a statistical authentication method to match the character strings by comparing the input time between each character from the initial input data. As shown in Table 5, of the two authentication methods using statistical safety, the preference authentication method is superior to the keystroke authentication method with regard to user convenience and the possibility of implementation, if the initial number of input characters and the required number of characters for authentication are examined.

In the calculation of the initial number of input characters, each question of the initial preference symbol selection question set shows an input of three characters. Only 90 characters are required and the initial questionnaires are extended to ninety questionnaires. In a typical situation, the required number of characters for authentication is only six: it is similar in convenience to the personal identification number (PIN) input method.

The preference symbols authentication method, which does not use a fixed match rate, involves a gradually increasing match rate whenever it is used, and can thus be considered superior to keystroke authentication.

For security verification, we examine the performance of our proposed system against shoulder surfing, man-in-the-middle, and dictionary attacks. The design algorithm includes a variable password system, whereby the password is changed at each instance of connection. Thus, our system guarantees security against shoulder surfing attacks. If the number of data values in the information storage is estimated to be 20 and six questions are extracted, the possible combinations are:

$$_{20}C_6 = \frac{20!}{(20-6)! \times 6!} = 38760 \tag{1}$$

If the number of initial data values is estimated to be 30, then

$$ {}_{30}C_6 = \frac{30!}{(30-6)! \times 6!} = 593775 \tag{2}$$

cases are calculated . If the appearance of each question at a particular position in a sequence is considered in the multiple choice from among six candidates, the cases are as follows:

$$ {}_{20}C_6 = \frac{20!}{(20-6)! \times 6!} = 38760 \times 6 = 232560 \tag{3}$$

$$ {}_{30}C_6 = \frac{30!}{(30-6)! \times 6!} = 593775 \times 6 = 3562650 \tag{4}$$

Hence, our system is secure against shoulder-surfing attacks.

Against man-in-the-middle attacks, intercepting information sent from the server in a terminal transmission is does not render the system insecure because authentication is achieved by the identifier stored in the server. To prevent an interception attack on information sent from a terminal in a server transmission, the transmitting data is sent using qualified public key encryption technique.

List attacks are blocked from the very beginning because our proposed system is a variable password system that is not affected by other passwords a user might have. The period of extracted questions is affected by the number of initially collected questions. If the number of initial questions is estimated to be 30 and the password used at a previous connection attempt is taken into consideration, the probability of estimating the next password is $\frac{1}{593775}$. Thus, we can conclude that our system protects against dictionary attacks.

Because our system improves a user's coincidence probability by restoring the incorrect answers, a safer system is formed by increasing the coincidence rate at the time of bulk connections. The dictionary attack can be avoided by disconnecting when an incorrect answer is provided more than three times. However, even if a connection is achieved through bypassing, the success probability of an attack, given the coincidence condition of correct answers to four questions of the basic six provided, is $\frac{1}{1296}$. The success rate of the dictionary attack against our authentication using five questionnaires, given the coincidence condition at an increased coincidence rate by restoring the incorrect answers, is $\frac{1}{46656}$: it thus provides superior security compared to a general PIN method.

5 Conclusion

The concept of attacks from within information systems cannot be understood by only considering external factors. The loss of information and intellectual property due to errors in the system, insider attacks, or even carelessness can be more serious than those caused by external attacks. The DRM method decides the range of accessibility of a given document according to its security level by differentiating the use authority of internal documents and data. This inevitably requires two-step or three-step authentication processes.

This paper addressed the disadvantage of existing password-based security systems, which cannot satisfy the conflicting demands of user convenience and security for data. The password system, which requires memorization within the range of human cognitive ability, has its limits. To solve this problem, numerous authentication methods are being researched, but no algorithm can simultaneously satisfy the competing demands for user convenience and low cost of infrastructure implementation. Our proposed authentication algorithm, which uses preference symbols, excels with regard to user convenience and cheap implementation.

As is evident from the statistical data, the general human recognition rate of preference symbols shows approximately 75% coincidence. This implies that more than four questions of six coincide. If the preference symbols to set the initial authentication data are carefully selected, they are expected to show superior coincidence rates than the average values.

The stored initial data generates six questions using a random question generator, sends them to a user terminal, and sends only the questions selected by the user to the identifier. The data that is sent, which uses a proven public key algorithm, is almost identical to the six-digit number. Thus, the traffic due to the connection can be said to be fairly light. All subsequent processing occurs in the server, and thus is safe against attacks.

Our proposed method carries out user authentication using only the initial settings and does not require memorization by the user; it is thus excellent in that it does not tax the user's memory. Improper password leaks are impossible in our system to begin with. It also provides security against external and internal system penetration attacks. Password leakage can also be prevented by cutting off the information storage and thereby negating the possibility of document exposure by the intervention of the administrator. In the future, considerable advancement in document security and second-step authentication is expected, if research on insider connection control, document security class setting, and searching coincidence rates of preference symbols continues to progress.

References

[1] T. Buch, A. Cotoranu, E. Jeskey, F. Tihon, and M. Villani. An enhanced keystroke biometric system and associated studies. In *Proc. of Student-Faculty Research Day (SFRD'08), New York, U.S.A*, pages C4.1–C4.7, May 2008.

[2] I. Buck, N. Govindaraju, J. Kruger, A. Lefohn, T. Purcell, and C. Woolley. Gpgpu: General-purpose computation on graphics hardware. In *Proc. of the 32nd International Conference on Computer Graphics and Iteractive Techniques (SIGGRAPH'05), Los Angeles, USA*, pages 1–1. ACM, August 2005.

[3] R. Guidorizzi. Integrated cyber analysis system. `http://www.darpa.mil`, November 2011.

[4] J.-H. Kang, J. Y. Kim, and E.-G. Kim. A study of the variable password generation method in internet authentication systems. *Journal of the Korea Academia-Industrial cooperation Society*, 14(3):1409–1415, March 2013.

[5] K. S. Killourhy and R. A. Maxion. Comparing anomaly-detection algorithms for keystroke dynamics. In *Proc. of the IEEE/IFIP International Conference on Dependable Systems and Networks (DSN'09), Lisbon, Portugal*, pages 125–134. IEEE, June–July 2009.

[6] Q. Liu, R. Safavi-Naini, and N. P. Sheppard. Digital rights management for content distribution. In *Proc. of the Australasian information security workshop (ACSW Frontiers'03), Adelaide, South Australia*, pages 49–58. ACM, February 2003.

[7] S. Liu and R. Kuhn. A study of the variable password generation method in internet authentication systems. *IEEE IT Professional*, 12(2):10–13, March-April 2010.

[8] A. Volchkov. Server-based computing opportunities. *IEEE IT Professional*, 4(2):18–23, March-April 2002.

[9] E. Yu and S. Cho. Keystroke dynamics identity verification-its problems and practical solutions. *Journal of Computers and Security*, 23(5):428–440, July 2004.

Author Biography

KwangCheol Rim received the PhD. Degrees in mathematics from Chosun University in 2006. He is currently a researcher in Information Technologies and information security. His current research interests are quantum cryptography, endpoint security and big data.