





Context Data Preprocessing for Context-Aware Smartphone Authentication

Sangjin Nam¹ , Suntae Kim¹ , Jung-Hoon Shin¹,
Jeong Ah Kim², and Sooyong Park³

¹ Department of Software Engineering, CAIT, Chonbuk National University,
567 Baekje-daero, Deokjin-gu, Jeonju-si, Jeollabuk-do, Republic of Korea
{potter930, stkim, shinjh}@jbnu.ac.kr

² Department of Computer Education, Catholic Kwandong University,
Beomil-ro 579 beon-gil, Kangneung-Si, Kangwon-Do, Republic of Korea
clara@cku.ac.kr

³ Department of Computer Science and Engineering, Sogang University,
35 Baekbeom-ro, Mapo-gu, Seoul, Republic of Korea
syPark@sogang.ac.kr

Abstract. This paper proposes an approach to carrying out context data preprocessing gathered from smartphone users to support context-aware authentication. Context-aware authentication is a technique to implicitly authenticate a smartphone user using contextual data (e.g., call log, location) without explicitly requesting the user's any actions. In order to enable context-aware authentication, a user's contextual data should be carefully processed for learning user's past contextual patterns in consideration of user's hourly, daily, weekly or monthly behaviors. In this paper, we gathered contextual data from 200 voluntary smartphone users for about 2 years and showed what the appropriate contextual data is preprocessing for performing context-aware authentication.

1 Introduction

Recently, most of the people use diverse mobile devices in their daily life. Among them, a smartphone is considered as the most popular devices because it basically can be used as a communication tool, and also people can use diverse complementary services in the single device [4, 5]. The growth of the use of the smartphone has caused an increase of the demand of user authentication techniques in the smartphone applications, as it has its own authentication methods and steps [3].

Most of the applications in a smartphone apply the id/password scheme for user authentication. However, the *id/password* scheme has a big and broad issue that it relies on the human's memory so that it is easy to forget. Some of the applications that handle secure data (e.g., bank or stock trading accounts) use a *digital certificate* issued by the public certificate authority or the *OTP* (One-Time-Password) scheme for user authentication. However, the digital certificate scheme forces a user to issue the certificate from the central authority with a complicated certificate issue process, and the *OTP* scheme needs extra devices and network bandwidth to get the one-time-password [1, 2, 10]. In addition to these, the diverse *biometric methods* [6] based on sensors of

the smartphone have been proposed, but it contains another issue regarding the low accuracy of the sensors and negative effects of the environmental factors such as illumination, humidity, etc.).

In order to address the issues, the context-aware authentication, so-called implicit authentication, techniques based on user's contextual data have recently been studied [6–9]. This technique uses call history record or location data that can be easily collected from a smartphone, carrying out the user authentication by comparing the past historical context data to the recent delta t time context data. However, the performance of the authentication technique highly relies on the preprocessing of the context data to appropriately characterize the user's behaviors.

This paper proposes the context data preprocessing technique for context-aware smartphone authentication. It consists of three steps. First, we introduce several types of context data that can be gathered from the smartphone and discuss its characteristics. Then, the gathered data is preprocessed in the three steps: (1) unification of the time unit for serializing the context data, (2) location data preprocessing, and (3) n-dimensional aggregation of the context data. The last step is to measure the quality of the preprocessed context data by applying statistical techniques. After the three-step preprocessing, diverse context-aware authentication schemes [6–9] may achieve higher performance. As the evaluation, we collected 200 voluntary smartphone users for about 2 years and showed the result of the sensitivity analysis with several data preprocessing parameters.

The rest of the paper is organized as follows: Sect. 2 introduces a representative authentication technique and discuss its pros and cons. Section 3 presents context data preprocessing technique composing of three major steps. Section 4 describes how we obtained the high-quality context-aware data obtained from the voluntary users with tuning the parameters. Section 5 concludes this paper.

2 Related Work

This section introduces several traditional authentication techniques and discusses its pros and cons. Also, it presents background on the context-aware authentication and its data preprocessing approaches.

2.1 Authentication Techniques

In this section, we introduce several traditional authentication approaches such as ID/Password, OTP, digital certificate and biometric technique and discuss its pros and cons. The ID/Password is the most popular and broadly used user authentication scheme not only in the smartphone but also in the general computers. This scheme is fast and relatively easy because a user just enters the password matched with his/her id as shown in Fig. 1(a). However, the scheme can work only if the user should memorize the pair of information. Because of this, most users tend to memorize several pair of id/password, otherwise, they consistently use a single id/password throughout applications. Furthermore, the application like Google Chrome provides a feature that keeps a user's id/password of a specific web site, which causes another security breach of the system. Additionally, in case of that a user forgets the pair of information, the user should carry out many tedious steps in order to recover the information [10, 13].

The OTP (One-time password) scheme is an authentication technique where a password is valid during only one login session, which is generally used as the complementary authentication. Figure 1(b) presents the steps of OTP. Once a user in OTP requests a one-time password to the application, the application generates a one-time password, and sends it back to the user. If the password that the user enters and that generated by the application is the same, the user is valid. Although its security level seems to be high thanks to the temporal password, a user should have an extra device or extra system (e.g., email) to obtain the password. Thus, it should rely on the other authentication scheme for the extra device or system accordingly. And also, it should spend extra cost to obtain the temporal password [11].

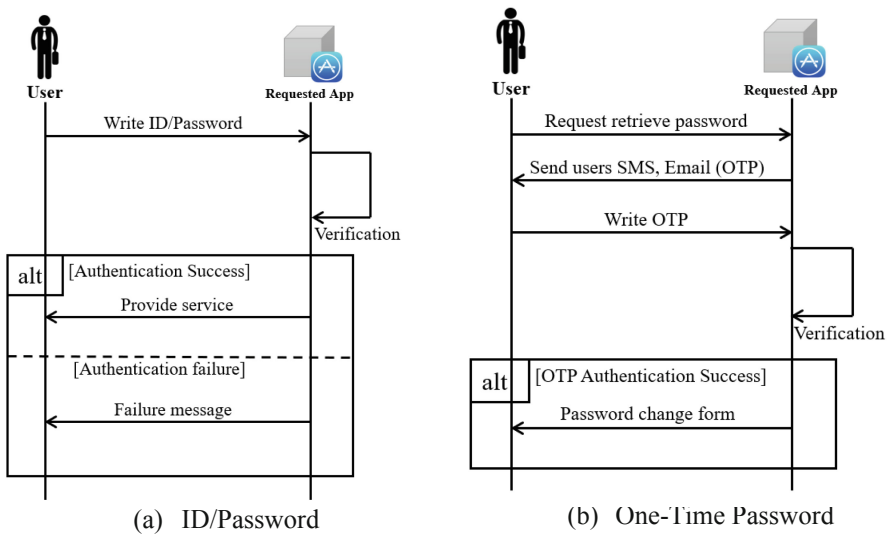


Fig. 1. The ID/password and OTP authentication scheme semantics

In addition to the two authentication scheme, the digital certificate is considered as a powerful authentication technique that is broadly used in the financial industry [21]. This technique depends on the certificate authority that generates the user’s pair of keys. As shown in Fig. 2, the user should submit user’s document that can guarantee his/her identity to an organization connected to the certificate authority and carry out the several steps to get the user’s pair of keys issued by the certificate authority. A user should regularly update the digital certificate, as it is only valid during a certain period of time. Once the user keeps the digital certificate issued by the certificate authority, he/she provides the certificate to the application with its corresponding password, and then the application requests the authority to validate the certificate and the password. This digital certificate authentication scheme contains several issues: (1) the digital certificate is usually stored in the local storage and is likely to be stolen; (2) it must use a specific application (e.g., a web browser); and, (3) a user has to perform several steps in offline to issue the digital certificate.

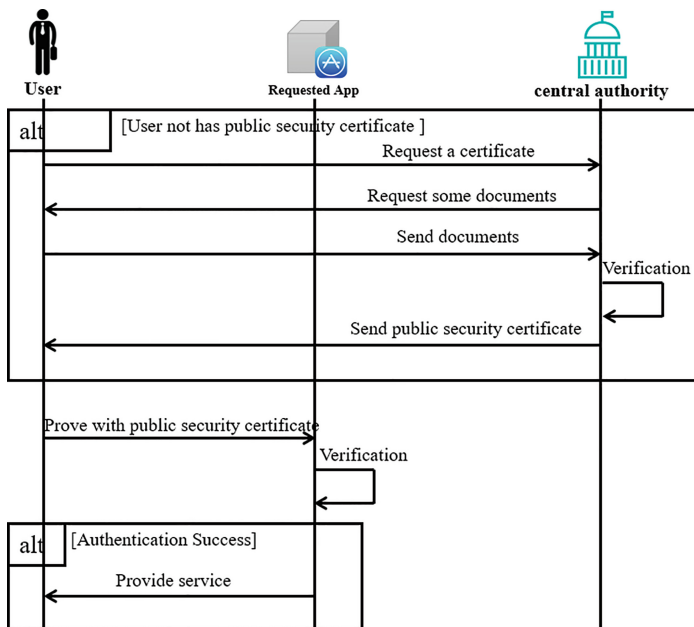


Fig. 2. The digital certificate scheme semantics

Recently, the biometric approach starts to be broadly used due to its convenience. The approach first stores the biometric information (e.g., fingerprint, iris, or face) [12–14] and authenticate a user with the pattern matching technology between the stored and the entered. The iris recognition approach [15] is a biometric identification that uses pattern recognition of each human's iris. Also, the face recognition approach has similar steps for the user's authentication, but the recognition success rate is not very high because the glasses and lighting have an heavy influence on the recognition. The fingerprint approach needs an extra device to capture the user's fingerprint, and its recognition rate depends on the humidity and hands foreign matters.

2.2 Context-Aware Authentication

The context-aware authentication, so-called implicit authentication, techniques based on user's contextual data have recently been studied [6–9]. This technique uses call history record or location data that can be easily collected from a smartphone, carrying out the user authentication by comparing the past historical context data to the recent delta t time context data. Figure 3 presents the overall process of context-aware authentication. First, a user just uses his/her own device (e.g., smartphone) that has several sensors. The device collects diverse user's contextual data (e.g., *CDR (Call Data Record)*, *Location*, *App Usages*). Second, the contextual data is processed in several ways and stored in the *context database*. Third, the user who wants to access specific application requests it to authenticate him/herself. The application compares the user's recent context to that of the context database using *Authentication Model*. If two data sets have a huge gap, the application denies the user's access, otherwise, it allows it.

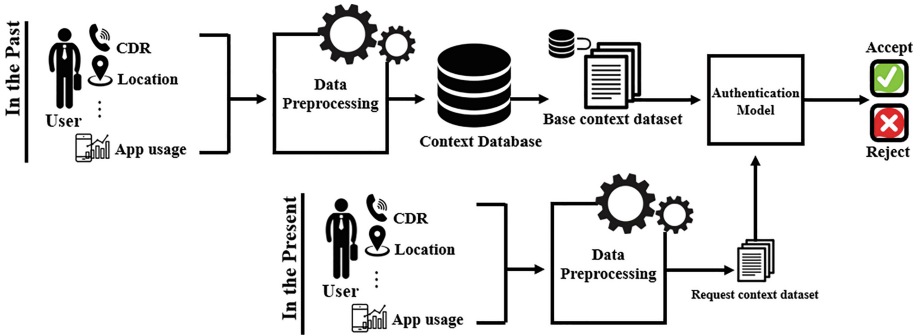


Fig. 3. The overall process of the context-aware authentication

The context-aware authentication assumes that a human has similar behaviors on the regular basis. Thus, if a user recently behaves differently than usual, it indicates that it needs to suspect that the user is the same person. For example, the user A wakes up at 7 am, goes to the company at 8 am and usually have a phone call with his/her parents during the commute time. In this case, the user’s location and call data in a specific time span can be considered as a context data. If the context data is different from that of the user, it is statistically suspected as a stranger.

Shi *et al.* [7] carried out the data preprocessing with context data aggregation based on the time-of-day, day-of-week, and separated the data into two groups: *good events* and *bad events*. The good event indicates the system event (e.g., call or location) that is already known. For example, the *incoming* or *outgoing* call number is one of the contact numbers stored in the smartphone. For the location, if the current location is the location that the user frequently stayed before, it is considered as a good event. Based on the good/bad event aggregation, the paper computes the authentication score as the following equations:

- $Score_{Positive} = \prod_i^n p(\Delta Time\ of\ Good\ Event \mid Time\ of\ day)_i$
- $Score_{negative} = \prod_i^n p(\Delta Time\ of\ Bad\ Event \mid Time\ of\ day)_i$
- $Score = Score_{GoodEvent} - Score_{BadEvent}$

In this approach, the data preprocessing indicates data aggregation or density based on the time span (e.g., time-of-day, day-of-week). It only considers the gap between two sets of events in a specific time span without considering the importance of type of event (e.g., call data vs. location).

Kayacik *et al.* [9] suggested similar context-aware authentication techniques. They established the authentication model with the special model and temporal model. In order to support the model, they created the context database with sensor data in associated with time as well as location. All sensor data are first grouped in terms of the specific time span and location, and then the data is statistically scored by the probability density functions (pdf) and conditional probability as below:

- $Score = \frac{\sum_i^n pdf(degree\ of\ sensor \mid Time\ of\ Day)_i}{n}$

3 Context Data Preprocessing Technique

This section proposes the data preprocessing technique for the context-aware authentication, composing of three steps: (1) data collection, (2) context data preprocessing, and (3) statistical quality evaluation of the data. From the following subsections, we present a detailed explanation.

3.1 Data Collection

Data collection is the first step of the context-aware authentication in the smartphone. The smartphone has several sensors and useful information that can be used as a user's context. It may be impossible to guarantee individual identity with just a single context data, but a combination of several context data may increase the precision of the authentication. The possible context data that can be used for the context-aware authentication is summarized as below:

- *CDR (Call Data Record)* is a set of incoming and outgoing calls in the smartphone. In the time of a day or a week, the user tends to make phone calls to similar people on the regular basis. Because of this, CDR is one of the possible context data. CDR data is composing of *types of calls* such as incoming and outgoing, *timestamp*, *receiver number*, *receiver name*, and *call duration*.
- *Location Data* indicates a user's location composing of a *latitude* and *longitude* information collected by GPS (Global Positioning System). It is considered as the most appropriate data for the context-aware authentication [7]. The location data consists of the latitude, longitude, altitude by the timestamp. Thus, it can be used to extract user's location moving patterns, the major spot where the user stays and how long.
- *SMS (Simple Text Message)* is another candidate. The data is composed of not only text message itself, but also the received time. Sometimes, this data contains very important contents for the authentication (e.g., authentication key issued by the authentication authority). However, SMS tends to contain diverse spam messages. Thus, the outgoing text message is generally used for authentication.
- *App Usage* indicates a history of a user's application usage. Like CDR and location data, a user tends to use similar applications depending on the location and time so that it is very useful information for context-aware authentication.
- *Typing Pattern* denotes a user's key input pattern of the smartphone virtual keyboard. Depending on the user, the key input speed, the typo frequency is different. Some of the research used to use this data for user authentication [13].

3.2 Context Data Preprocessing

This subsection describes how the context data is processed for context-aware authentication. Among the candidate data set mentioned in the previous subsection, we selected the CDR and location data as the key data set for the context-aware authentication. This is because those are the most representative and fundamental data that most of the smartphone support. In addition, the literature [7] showed the two datasets

characterizes the user’s context very well. The context data preprocessing consists of three steps as presented in Fig. 4. First, it starts with the unification of the timestamp for the diverse context data. Then, the location data is specially processed, because the longitude and latitude are quite fine-grained, it is inappropriate to use context-aware authentication. Based on the data, the hierarchical aggregation is performed depending on the appropriate time span.

3.2.1 Unification of the Time Unit

The first step is the unification of the time unit, aiming at the serialization of all contextual data at the end of this step. Depending on the context data from different smartphones, the time data format is very different. For example, some of the data has the time format ‘yyyy-MM-dd HH:mm:ss.SSS’ and other has a Unix timestamp composed of 10 digit numbers. Thus, it is inevitable to make the time format the same for serializing all system events depending on its occurring time.

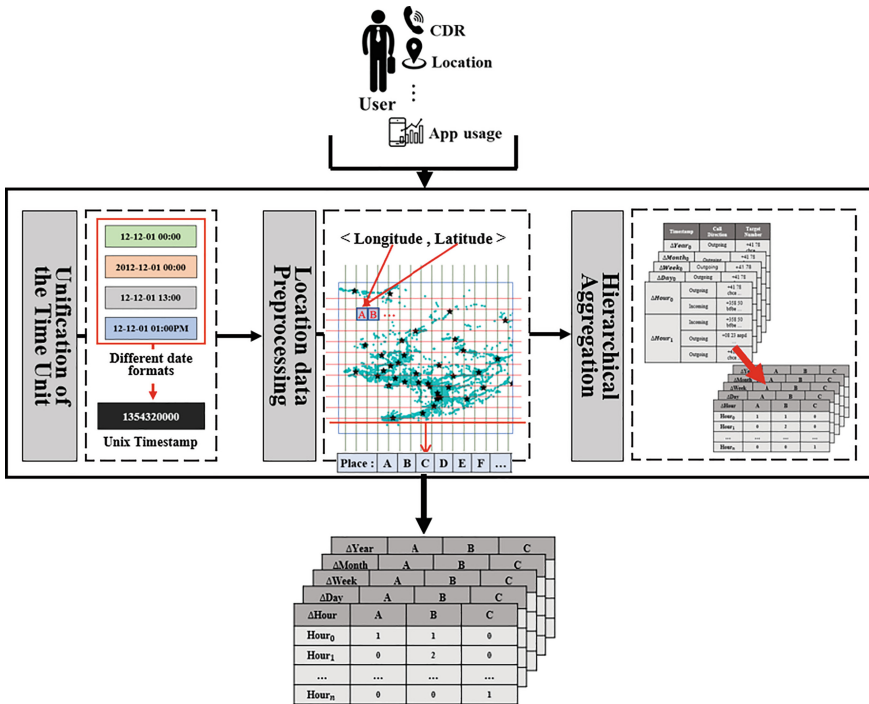


Fig. 4. The steps of the data preprocessing for context-aware authentication

3.2.2 Location Data Preprocessing

Location Data indicates a user’s location composing of a *longitude* and *latitude* information collected by GPS. By using the location data, we can extract the user’s location moving patterns and major spot where the user stays and how long he/she does

in the stop. However, as the longitude and latitude is very small number composed of 10^{-10} (e.g., 46.5212053706, 6.6190893676), it is not efficient to use them to compare two locations directly. There might be several possible approaches to preprocessing location data. One of them is applying the clustering approaches of machine learning such as K-Means [16] or agglomerative hierarchical clustering [17]. However, it is a hard problem to decide the appropriate cluster number, and also the clustering approach loses too much information for the location.

Another possible approach is to make a grid map based on the latitude and longitude as shown in Fig. 5. Depending on the α that indicates the number of cells of the entire world, the place is simply computed like the equation in the figure. Thus, we need to decide the appropriate α . In case of the $\alpha = 10^7$, 0.001×0.001 of the latitude and longitude gap indicates the $100 \text{ m} \times 80 \text{ m}$ size cell in the real world. In the evaluation section, the α is obtained by the sensitivity analysis.

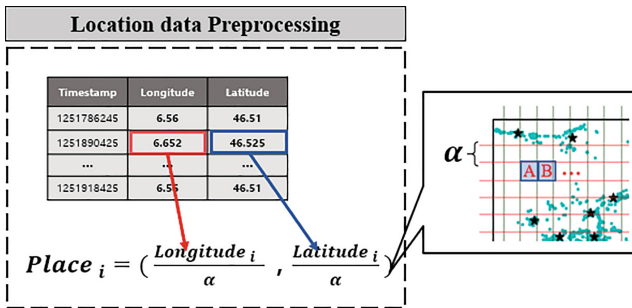
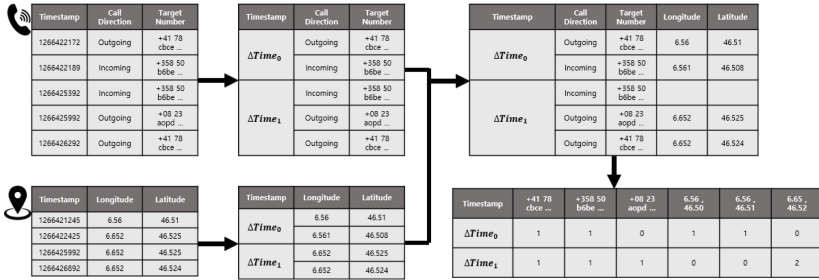


Fig. 5. Location data preprocessing

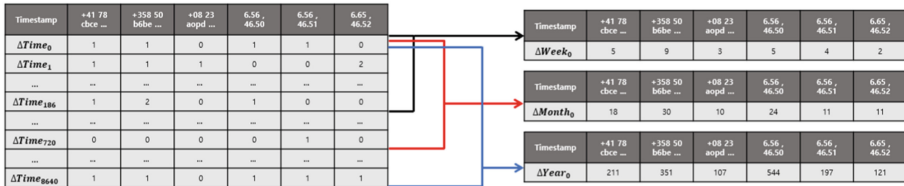
3.2.3 Hierarchical Aggregation

This step is a hierarchical aggregation of the serialized context data with grid-mapped location data. It starts with making a feature vector composed of context data that occurs during $\Delta Time$. Figure 6(a) denotes the aggregation steps of the CDR and location context data occurred during the fundamental time span (e.g., 1 h or 2 h). In the figure, the left tables are the raw CDR and location data, and then those are grouped into the $\Delta Time$ base (see the two tables in the center column). Then, the two data are merged into one table, and each data are aggregated and placed at each column. Thus, the one row of $\Delta Time$ denotes the number of context data that occurred during the time span. It should be noted that $\Delta Time$ is considered as the minimal time of gathering the context data for requesting authentication.

The next step is a hierarchical aggregation from the fundamental time span aggregation. If the fundamental time span is 1 h, the hierarchical aggregation time span might be 1 day or 2 days, 1 week or 1 month as shown in Fig. 6(b). The levels of the hierarchy is decided to the experiment, however, we can imagine that the $\Delta Time$ and the levels of hierarchical aggregation should characterize a human's life pattern.



(a) Fundamental Time Span Aggregation



(b) Hierarchical Aggregation

Fig. 6. Hierarchical aggregation steps

3.3 Statistical Quality Evaluation of the Data

After aggregating context data, the data quality should be evaluated. As the context-aware authentication is an approach to comparing the recent context data to the past context data, the recent context data indicates the context data gathered during $\Delta Time$ and the past context data can be the past data gathered during $\Delta Time$ and the result of hierarchical aggregation. The high-quality data denotes that the recent and past context data should have the same mean and variance for the same user. Otherwise, the mean and variance are different for the different user. Thus, the α and $\Delta Time$ should be decided to make high-quality data.

Statistically, *Two-Sample T-Test* is the approach to statistically check if the means of two sets of data is the same or not. Thus, the *p-value* greater than 0.05 of the T-Test denotes that the means of the two data sets statistically are the same [18]. For checking if the distribution (i.e., variance) of the two sets of data, we can apply F-Test where we can consider that distribution of two sets of data is statistically the same if the *p-value* is greater than 0.05 [19]. We can guess that the recent smartphone user is the same user with the past if the *p-value* resulting from T-Test and F-Test of the recent context data and the past context data is greater than 0.05. Otherwise, the recent and past users are different. To make the two tests passed for the same user’s context data, we should decide the parameters α and $\Delta Time$.

4 Experiment: Real-World Context Data

In this section, we present an experiment on the data preprocessing for the real-world context data. Thus, we first introduce the context data set and then present how we obtain the best parameters to have high-quality context data using the statistical approach. Finally, we discuss the result in the last subsection.

4.1 Experimental Setting

We have applied our approach to the MDC (Mobile Data Challenge) dataset [20], which is a collection of smartphone data such as accelerometer, network connections, calendar, CDR as well as GPS gathered from about 200 users for two years from 2012. We extracted the five user's context data, because only small number of the user's data has complete and location data during the period. Table 1 presents the data set.

Table 1. The number of the five smartphone user's dataset

User ID	Number of CDR	Number of GPS	Number of Data
5938	7	1931	1938
5973	3	1717	1720
5928	6	922	928
5976	32	814	846
6177	10	797	807
...

4.2 Experiment Result

We carried out the MDC data preprocessing according to the aforementioned steps: (1) unification of the timestamp, (2) location data preprocessing, and (3) hierarchical aggregation. Then, we tried to find the best parameters such as α and the $\Delta Time$ by

Table 2. Experiment result

α	Request data	base data	F-test p-value	Average of F-test p-value	T-test p-value	Average of T-test p-value	
10^8	1hour	1 hour before	3.6735E-05	0.033345578	0.572925	0.456889815	
	2hour	2 hour before	0.1		0.6197		
	1day	1 day before	1.4798E-267		0.178044444		
	1hour	the same hour before 1 week	0.3719	0.460938095	0.879433333		
	2hour	the same hour before 1 week	3.96E-01		9.70E-01		
	1day	The same weekday of last	0.615414286		0.9787		
	1week	last week	0.113677778	-	0.934966667		-
	1month	last month	2.37E-02	-	9.16E-01		-
10^7	1hour	1 hour before	0.036021898	0.017174757	0.87131	0.702263056	
	2hour	2 hour before	0.015502372		0.8531125		
	1day	1 day before	6.5065E-171		0.382366667		
	1hour	the same hour before 1 week	1.63158E-09	0.107085715	0.704514286		
	2hour	the same hour before 1 week	0.321257143		0.962828571		
	1day	The same weekday of last	1.10893E-44		0.278175		
	1week	last week	0.635525	-	0.9754625		-
	1month	last month	0.0935	-	0.912333333		-

applying the F-Test and T-Test statistical techniques. We performed the experiment with two $\alpha \times 10^7$ and 10^8 , the five $\Delta Time$ 1 h, 2 h, 1 day, 2 days, 1 week and 1 month. Also, we compared the request data (recent data) to the diverse past data (base data) of the same user. Table 2 summarized the result.

The first row of Table 2 can be understood as the follows: (1) the location separated by the $100\text{ m} \times 80\text{ m}$ grid ($\alpha = 10^8$), (2) we compared the two data sets, each of which is the recent 1 h data and the 1 h before the 1 h, (3) the variance of the two sets is different and the mean is the same, because the p-value of F-test is less than 0.05, and the p-value is greater than 0.05. In order to make the data valid for the context-aware authentication, most of the p-value of F-Test and T-Test should be greater than 0.05. Thus, we can conclude that comparison between 1 h – 1 h before is not appropriate though the two means are the same.

We highlighted the data set which has the p-value less than 0.05 with the bright red color in the table. According to the result, most of the means in the comparison is the same (see the result of T-Test) and the variance (i.e., distribution) is different. Also, the result of F-Test in case of $\alpha = 10^7$ indicates the comparison has a different distribution, though the two data set is from the same user. Also, we can recognize a comparison between 1, 2 h, 1 day and the same hour before 1 week and the same weekday of the last provides better data quality.

According to the experiment result, we can conclude that the $100\text{ m} \times 80\text{ m}$ grid map of the location provides better performance, and aggregating the data depending on the week, and comparison data such as 1 h, 2 h to that before 1 week showed better performance. Thus, we can conclude that the hierarchical aggregation should be 1 h, 1 day, 1 week, 1 month, and 1 year as shown in Fig. 7.

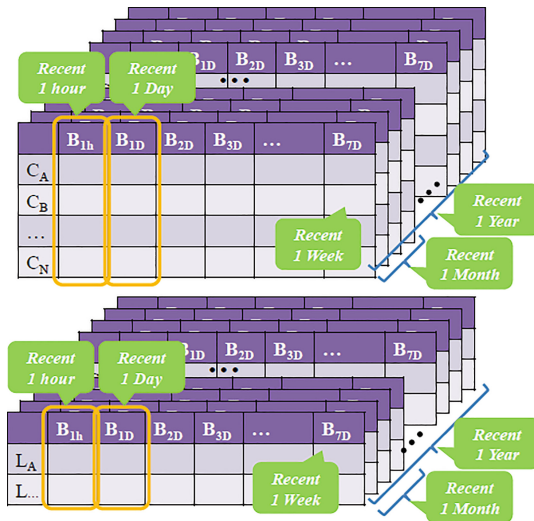


Fig. 7. Appropriate hierarchical aggregations

5 Conclusion

In this paper, we proposed the data preprocessing approach for context-aware authentication in a smartphone. For the authentication, we first summarized the candidate context data that can be gathered from the smartphone. Then, we proposed the three-step data preprocessing approach consisting of (1) unification of the time unit for serializing the context data, (2) location data preprocessing, and (3) n-dimensional aggregation of the context data. For the evaluation, we applied our approach to the MDC dataset and showed how we obtained the best parameters by using F-Test and two sample T-Test. As future work, we have a plan to enhance our approach more so that we will establish the model of context-aware authentication and implement it.

References

1. Herley, C.: So long, and no thanks for the externalities: the rational rejection of security advice by users. In: Proceedings of SACMAT (2009)
2. Hulsebosch, J.R., Salden, H.A., Bargh, S.M., Ebben, P.W.G., Reitsma, J.: Context sensitive access control. In: Proceedings of SACMAT (2005)
3. Nachenberg, C.: A window into mobile device security: examining the security approaches employ. SSR (2011)
4. Whitney, L.: Smartphones to dominate PCs in Gartner forecast. CNET Business Tech News (2010)
5. Rainie, L., Anderson, J.: The Future of the Internet III. Pew Internet Project (2008)
6. Zhang, F., Kondoro, A., Muftic, S.: Location-based authentication and authorization using smartphone. TrustCom (2012)
7. Shi, E., Niu, Y., Jakobsson, M., Chow, R.: Implicit authentication through learning user behavior. In: Burmester, M., Tsudik, G., Magliveras, S., Ilić, I. (eds.) ISC 2010. LNCS, vol. 6531, pp. 99–113. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-18178-8_9
8. Markus, J., Shi, E., Golle, P., Chow, R.: Implicit authentication for mobile devices. USENIX (2009)
9. Hilmi, G.K., Just, M., Baillie, L., Aspinall, D., Micallef, N.: Data driven authentication: on the effectiveness of user behaviour modelling with mobile device sensors. MoST (2014)
10. Lamport, L.: Password authentication with insecure communication. Commun. ACM **24** (11), 770–772 (1981)
11. Huang, C.-Y., Ma, S.-P., Chen, K.-T.: Using one-time passwords to prevent password phishing attacks. J. Netw. Comput. Appl. **34**(4), 1292–1301 (2011)
12. Xi, K., et al.: A fingerprint based bio-cryptographic security protocol designed for client/server authentication in mobile computing environment. Secur. Commun. Netw. **4**(5), 487–499 (2011)
13. Nilesh, A., Salendra, P., et al.: A review of authentication methods. Int. J. Sci. Technol. Res. **5**(11), 246–249 (2016)
14. Bhatia, R.: Biometrics and face recognition techniques. Int. J. Adv. Res. Comput. Sci. Softw. Eng. **3**, 93–96 (2013)
15. Kak, N., Gupta, R.: Iris recognition system. Int. J. Adv. Comput. Sci. Appl. **1**, 34–40 (2010)
16. Hartigan, J.A., Wong, M.A.: A k-means clustering algorithm. J. Roy. Stat. Soc. Ser. C (Appl. Stat.) **28**(1), 100–108 (1979)

17. William, H.E.D., Edelsbrunner, H.: Efficient algorithms for agglomerative hierarchical clustering methods. *J. Classif.* **1**, 7–24 (1984)
18. Snedecor, G.W., Cochran, W.G.: *Statistical Methods*. 8th edn. Iowa state University Press (1989)
19. Shen, Q., Faraway, J.: An f test for linear models with functional responses. *Statistica Sinica* **14**, 1239–1257 (2004)
20. Mobile Data Challenge (MDC) Dataset: Dataset Distribution Portal. <https://www.idiap.ch/dataset/mdc>. Accessed 6 Feb 2019
21. Perlman, R.: An overview of PKI trust models. *IEEE Netw.* **13**(6), 38–43 (1999)