Application of Keystroke Dynamics Modelling Techniques to Strengthen the User Identification in the Context of E-commerce

Germán M. Concilio, Enrique P. Calot,
Jorge S. Ierache, and Hernán D. Merlino

Laboratorio de Sistemas de Información Avanzados,
Facultad de Ingeniería, Universidad de Buenos Aires,
Buenos Aires, Argentina
{gconcilio,ecalot,jierache,hmerlino}@lsia.fi.uba.ar

Abstract. Keystroke dynamics is a biometric technique to identify users based on analysing habitual rhythm patterns in their typing behaviour. In e-commerce, this technique brings benefits to both security and the analysis of patterns of consumer behaviour. This paper focuses on analysing the keystroke dynamics against an e-commerce site for personal identification. This paper is an empirical reinforcement of previous works, with data extracted from realistic conditions that are of most interest for the practical application of modelling keystroke dynamics in free texts. It was a collaborative work with one of the leading e-commerce companies in Latin America. Experimental results showed that it was possible to identify typists with an accuracy of 89% from a sampling of 300 randomly selected users just by reading comment field keystrokes.

Keywords: keystroke dynamics, biometrics, e-commerce, user identification, user classification

1 Introduction

The characteristics that help make a handwritten signature a unique human identifier also provide a unique digital signature in the form of a stream of latency periods between keystrokes. The handwritten signature has a parallel on the keyboard. The same neurophysiological factors that make a written signature unique are also exhibited in a user's typing pattern[12]. Biometric characteristics are unique to each person and have advantages as they cannot be stolen, lost, or forgotten[19,1,5].

During a user session, every key-press generates two events, one when the key is pressed and other on its release. A typing rhythm, consisting of the intervals between keys, albeit random at first sight and naturally variable, can provide meaningful information about the user like its handedness[18] or its emotional state[6]. What is more, timing data can be enough to identify the typist. The objective of keystroke dynamics analysis is to study biometric timing signatures and, which information can be inferred from them and their sources of noise.

Passwords and fixed short texts fit the general framework of keystroke dynamics well; being short sequences typed in a row and repeated often, their characteristic timings tend to be reasonably stable in a broad sense. The free text does not enjoy the same privileges. Typing errors, corrections, misspellings, interruptions, pauses to think and attention lapses which are impossible to predict, interfere with the source timing data. Furthermore, short-term variations due to daily tiredness, stress or emotional shifts and long-term effects of health and re-education of typing skills are strictly unavoidable. Due to the mentioned difficulties, the problem of free text analysis of keystroke dynamics has only recently begun to be attacked[11] and in slightly controlled conditions. In [9] the finite context modelling of keystroke dynamics in free text is examined, reporting promising results.

In this paper, we present new experimentations as an empirical reinforcement of the work done by Calot *et al.*[3,10]. Also, we use data extracted from realistic conditions that are of most interest for the practical application of modelling keystroke dynamics in free texts. Similarly to [10], we examine the performance of techniques for keystroke dynamics analysis in free texts under evaluation conditions which are much harsher than the used ones initially[11,2,9] but we extend it to the e-commerce domain.

According to [9], establishing the universality of results has proven very difficult in keystroke dynamics based models due to the inherent complexities of classifying this noisy behaviour. Replication and implementation efforts using different training and evaluation datasets have often lead to astonishingly varying error rates in comparison with the original study, both for static passwords[5,15] and free texts[17,13].

This paper is a collaborative work between the LSIA and one of the leading e-commerce companies in Latin America. We propose to analyse if it is possible to strengthen the `user_id` from the collection of time series data describing the dynamics of typing, or typing patterns for short. The technique of modelling by finite contexts[9] has been adapted for the specific context of data experimentation of an e-commerce platform. The company facilitated the employment of its dataset for our experimentation. The general framework we used admits many implementation parameters and selection strategies; their effects on classification performance were studied.

2 Description of the Problem

In the specific context of e-commerce, it is possible to generate an account matching model —or authentication problem—. It consists of comparing previous behaviours of a given user with the current behaviour to discover if it is the same person, i.e., it seeks to validate that the person who is interacting with the platform is who he is claiming to be. In this context, we present a classification problem about two mutually exclusive classes, namely false or positive, depending on whether the current session is considered valid for the user with whom it was previously authenticated based on the previous behaviour with the platform's points of contact.

Nevertheless, in the present paper, we propose to analyse the more complicated problem of user identification, less treated in previous works but of more significant interest for the forensic analysis. Identification is about finding the actual person with the same pattern over an extensive collection of typing patterns, while authentication is only checking against the assumed-right person. There are endless possibilities here, and some examples are: detecting fraudulent accounts (different `user_id`, same real person), detecting the identity of an intruder on a valid account (detecting the `user_id` of the intruder), or clustering compromised accounts; all by analysing keystroke dynamics on comments written on the platform.

A significant contribution of the present work is that we based it on an experimental sampling randomly selected without any bias, where users used their everyday keyboards resulting in entirely different settings for each of them. None of them were aware of this study, and, of course, the real identity of the person behind the user of the e-commerce platform was kept confidential, employing only the `user_id`. The device, web browser, operating system used, the configuration and type of keyboard may vary among different users and even for the same user between sessions. In addition, there is a temporal variability between sessions.

Keystrokes are fundamentally noisy; not only do we face the unpredictable and uncontrollable variations that occur due to physiological changes of the user —such as tiredness, emotional states, and consumption of medicines or other substances that may affect motor performance in a subtle way[6]—, but also we take information from an ecosystem of keyboards fixed, portable, tactile, among others; with different configurations and we also intend to integrate the modelling of keystroke dynamics in practical systems.

2.1 Free and Static Texts

The continuous authentication of users while typing employs the same structures and techniques as the verification of static keys (parameters, classifiers, distance metrics, characteristic vectors) although it adds some distinctive limitations. Among them, the need to fragment the input text to avoid noise originating from unnecessary pauses in data entry as effective measurements of parameters; and the impossibility of using metrics that take into account the correlation between keys since we expect the verified text to be substantially different from the training text[8]. In static texts such as passwords, the correlation of the types of events is very high —especially if the passwords do match— with a higher number of comparable events per amount of typed text. Unfortunately, as this work is on free texts, we are forced to use less accurate models to prevent the rate of false rejections from growing excessively.

Not all the suitable classifiers for static keys perform well with free texts, and in many cases, it is not even obvious how to adapt them to this situation.

Among the general purpose classifiers, those based on metric spaces have provided excellent results[7,16]. However, two new techniques of classification of specific purpose called metric R and metric A, initially proposed by Bergadano, Gunetti and Picardi in [2] and [11], stand out above the rest and have been used by additional authors[17] after optimising them for execution in real time, since the computational cost of these techniques is not irrelevant[10].

2.2 Theoretical Considerations and Common Techniques

The most commonly used characteristic parameters are the hold time —latency between key-press and release—, wait time —latency between key release and next key-press— and flight time —latency between successive key-press events—; to a lesser extent, average typing speed, probability of error or usage frequency of backspace and delete, key usage patterns and key release order when writing uppercase letters. Usually, the timing between consecutive keys —called digraphs— are used, but occasionally latencies of longer groups are chosen[2].

According to [10], probably every technique for classification and machine learning has been used for the analysis of keystroke dynamics, ranging from simple metric spaces —with norms such as Euclidean, Manhattan, Mahalanobis and an infinitude of variations and combinations, also approaches such as outlier counts— and k-NN to state-of-the-art classifiers such as support vector machines or random forests; artificial neural networks, fuzzy logic and genetic algorithms have been tried too, with not very promising results. Killourhy and Maxion's review[15] has become a classic; an updated one can be found at [14].

3 Experiment and Results

3.1 Dataset

In an e-commerce application, there are limitless points of contact connecting an individual and the purchase process where the keyboard is used enabling acquisition of keystroke dynamics while the user is ignoring that his typing rhythms are being analysed to improve his security.

Keystrokes extracted from the contact points related to user authentication (password entry, email and username) and checkout (credit card data) are useful regarding security as a second factor of authentication and are compatible schemes with verification of static keys. On the other hand, keystrokes extracted from forms —such as an inquiry of a buyer to a seller about a particular product and its response, the rating of a purchase from the buyer to the seller and vice versa, claims, among others— are of variable length and they have a considerably long text extension, which is why it is more appropriate to model them with keystroke dynamics techniques for free texts; and more appropriate for our analysis.

We acquired a dataset from a production environment and not in a laboratory setting. As a plus, we did not choose the subjects from a university population but a global population of e-commerce site users. Different keyboards with diverse regional configurations were used, even by the same individuals, and should be the ones the users have at home. Typing skills vary from mediocre to excellent. Noise factors affecting typing cadences such as interruptions, attention lapses and short-and-long-term emotional states, mood or health variations of the individuals were not excluded.

For each session, the timing of key down and key up events was recorded with a precision of 1 millisecond and the presumed user identity, previously authenticated with the valid credentials, is included. Due to the real world nature of the capture setup, we cannot guarantee that all sessions logged in with a particular identity are not used afterwards by a different user and, indeed, we suspect this is the case with at least some of them, but not enough to ruin the experiment. At worst, the reported results can be considered artificially degraded by this artefact, though not to a large extent[9].

The original dataset for experimentation provided by the company had 97,184 sessions of 5,673 users, many of the sessions of the original dataset did not have enough keystrokes to be considered for evaluation due to the length requirement. We accepted as valid sessions only the ones having at least 20 characters. The quantity minimum of valid sessions per user was set at 20. Since the task is unrestricted, some keystrokes might correspond to arrow keys and other special keys used for edition and navigation; in order not to negatively bias the results, the keystroke count referred to above considers only alphanumeric characters. Besides, all the sessions that have pressed the paste keys (ctrl+V or command+V) were discarded from the analysis, as they noticeably alter the user's natural keystroke behaviour.

After an adaptation and normalisation of the original dataset, to evaluate the proposed method, the final dataset consists of time series describing the keystroke dynamics from 372 users over eight months. Each user has in average 32 keystroke patterns (within 20–80), resulting in a collection of 11,766 keystroke patterns in total. Although all those cadences of less than 20 characters are filtered, the extension of the sessions shows a considerable variability reaching up to 300 key-pressures; the average extension for the entire data set is 56 keys per session. The total number of characters per user is approximately 1,650 characters (with a minimum of 1,000 and a maximum of 4,000). The intervals between successive sessions of the same user are also strongly variable, with values ranging from a few hours to several days.

3.2 Experimental Setup

We assume that the real person behind a `user_id` is always the same. We used various samples of randomly selected users to evaluate the behaviour of distinct identification methods with many customisable aspects. For each user, their keystrokes were subdivided into two mutually exclusive subsets, labelled A and B, previously applying a shuffle operation because the dataset was chronologically sorted. In this way, we generated two patterns with the same identity or `user_id`. For example, for the initial sample of 50 users, 100 patterns were obtained, two per user, one belonging to group A and another one to B. Through the modelling of the keystroke dynamics of each pattern we try to discover which other pattern is more likely to share the same `user_id`. We count one success when we identify the `user_id` of set A with the least pattern distance to the same `user_id` of set B.

3.3 Reporting Metrics

The simple metric used to grade the quality of the identifying method is the percentage of correctly matched users, that is the amount of `user_id` from the set A matching the same `user_id` in the set B over the total amount of users in the sample.

3.4 Implementation and Results

Specific tools were developed to test several configurations of the proposed method in experimental patterns in e-commerce.

We adopted the modelling by finite contexts, employing the keyboard event as a context of order N (differs from the implementation of González *et al.*[9] that utilises key-based contexts and models flight, wait and, hold time). Here, all the contexts are analysed without differentiating flight, hold and wait time of a specific key but using event-related contexts where the combination of action —press or release— and key are considered a unique event.

To find the optimal configuration of the identification algorithm, we randomly selected a sample of 200 users from a pool in the dataset. We varied the context sizes from 1 to 10. To calculate the distance between two users, we selected only the common intersection of events and built two vectors with mean and variance per user. Finally, we fed the distance algorithm with the mean vector for both users —normalised Euclidean additionally required the variance vectors—. We employed the following distance metrics: Metric A[11], Euclidean norm ($\mathcal{L}_2$), normalised Euclidean norm ($\bar{\mathcal{L}_2}$), Manhattan norm ($\mathcal{L}_1$), and Metric R[11]. Table 1(a) shows the experimental results, registering an advantage in effectiveness rates of Euclidean distance and the R metric over the other ones.

Due to computational power constraints, we used up to a maximum context order of 10; notwithstanding this limitation, Table 1(a) attests that the optimal classification performance peaks around context lengths in between four and seven events.

Table 1(b) presents the experimental results obtained by varying the sample size in steps of 50 from a range of 50–300, all with a fixed context of 5 events arbitrarily chosen. The same table, illustrated in Fig. 1, uncovers that the effectiveness rates achieved by all the adopted distances are stabilised starting from the 200-user samples.

3.5 Key Observations

It was found that considering timing data of certain special keys —including modifiers (shift, ctrl, alt), navigation (arrows, page up, page down) and correctors

Table 1. Experimental results in user identification for (a) different context orders and (b) different user sampling sizes

Order	A	$\mathcal{L}_2$	$\bar{\mathcal{L}_2}$	$\mathcal{L}_1$	R
1	77	84	74	80	81
2	83	83	80	83	85
3	83	86	78	82	86
4	83	85	77	82	87
5	83	85	76	81	87
6	82	85	76	82	86
7	82	85	77	81	86
8	81	85	77	81	85
9	81	84	78	81	85
10	79	84	77	81	84

(a) Different context orders with de sample of 200 users.

Users	A	$\mathcal{L}_2$	$\bar{\mathcal{L}_2}$	$\mathcal{L}_1$	R
50	88	96	92	94	90
100	82	90	85	87	90
150	74	85	78	80	86
200	83	85	76	81	87
250	80	86	78	79	88
300	78	86	77	80	89

(b) Different user sampling sizes with 5 as context order.

(backspace and delete)— worsened the classifier's performance even though their consideration in the contexts of the next keystrokes improved it.

4 Concluding Remarks

Despite all the noise, we obtained an effectiveness of between 96% (50 users sampling size) and 89% (300 users sampling size) in the identification of users. The experimental conditions detailed in Sec. 3.2 show that the times taken from the keystrokes typed in online purchases rating forms inside the e-commerce platform can be employed as a practical application method to strengthen the identification of users.

Table 1(a) reveals that the best classifiers are those that adopted the metric R in almost all cases. Still, the versatility and faster execution times of the classifiers based on the Euclidean distance makes it a very competitive alternative. Fig. 1 shows that the effectiveness rates achieved by all the adopted distances are stabilised starting from the 200-user samples.

Besides, it was established that the method proposed in [9] and replicated in [10] is still useful in the context of electronic commerce. The classification performance cannot compete with other biometric systems but has the advantage of being entirely non-intrusive, transparent, and not requiring additional hardware or user actions. The selected method seems promising, and there is still room for further improvements in this particular context.

4.1 Future Lines of Research

As future lines of work, we are planning to operate with other forms of model selection (the highest order context); expand the time lapse of the experiment

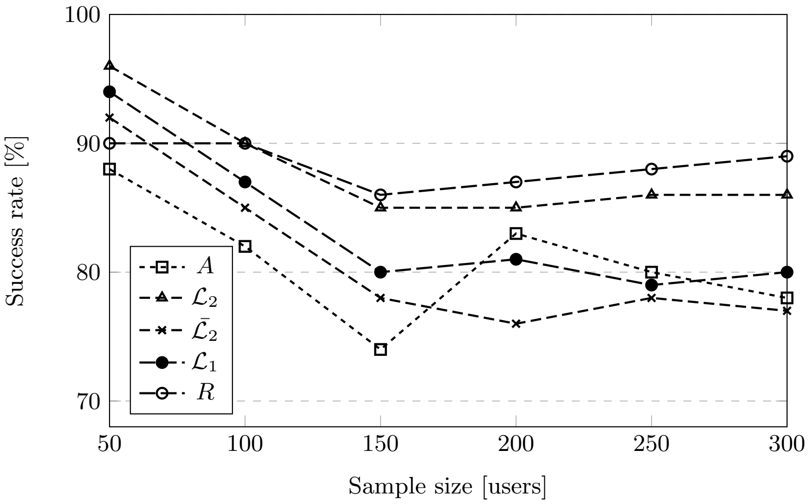

Fig. 1. Success percentage for samples from 50 to 300 for five metrics.

to make it last at least a year; utilise the algorithms to assert if the current user already existed on the platform (threshold at the minimum distance to be considered the same user); and optimise the code to be able to perform empirical tests with a higher number of users.

Our future lines of research will focus on eduction of emotions, initially contrasted against facial images and voice recognition, the results had already been explored in the UBACyT 2014–2017 GEF and will also be capitalised concerning the use of biometric EEG markers[4]; additionally, its applications in e-commerce. Also, the influence of emotional states and other noise sources to mitigate their adverse effects on classification.

References

1. Araujo, L.C.F., Sucupira, L.H.R.J., Lizarraga, M.G., Ling, L.L., Yabu-uti, J.B.T.: User authentication through typing biometrics features. Signal Processing, IEEE Transactions on 53(2), 851–855 (2005), `http://lsia.fi.uba.ar/papers/araujo05.pdf`
2. Bergadano, F., Gunetti, D., Picardi, C.: User authentication through keystroke dynamics. ACM Transactions on Information and System Security (TISSEC) 5(4), 367–397 (Nov 2002), `http://doi.acm.org/10.1145/581271.581272`
3. Calot, E.P.: Keystroke dynamics keypress latency dataset. Database (Jan 2015), `http://lsia.fi.uba.ar/pub/papers/kd-dataset/`
4. Calot, E.P., Ierache, J.S.: Multimodal biometric recording architecture for the exploitation of applications in the context of affective computing. In: Proceedings del XXIII Congreso Argentino de Ciencias de la Computación (La Plata, 2017). pp. 1030–1039. No. 10529 (2017)

5. Calot, E.P., Rodríguez, J.M., Ierache, J.S.: Improving versatility in keystroke dynamic systems. In: Proceedings del XIX Congreso Argentino de Ciencias de la Computación. No. 5606 (2013), http://ir.cs.uns.edu.ar/downloads/cacic_2013/11wsi.pdf
6. Epp, C., Lippold, M., Mandryk, R.L.: Identifying emotional states using keystroke dynamics. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 715–724. CHI '11, ACM, New York, NY, USA (2011), http://doi.acm.org/10.1145/1978942.1979046
7. Flior, E., Kowalski, K.: Continuous biometric user authentication in online examinations. 2014 11th International Conference on Information Technology: New Generations 0, 488–492 (Apr 2010), http://lsia.fi.uba.ar/papers/flior10.pdf
8. González, N.: Utilización de contextos finitos para el modelado de la dinámica de tecleo en esquemas de autenticación mixta (Apr 2014), http://lsia.fi.uba.ar/papers/gonzalez14.pdf, internal paper from Laboratorio de Sistemas de Información Avanzada, Facultad de Ingeniería, Universidad de Buenos Aires, Ciudad Autónoma de Buenos Aires, Argentina
9. González, N., Calot, E.P.: Finite context modeling of keystroke dynamics in free text. In: Biometrics Special Interest Group (BIOSIG), 2015 International Conference of the. pp. 1–5 (Sep 2015), http://ieeexplore.ieee.org/document/7314606/
10. González, N., Calot, E.P., Ierache, J.S.: A replication of two free text keystroke dynamics experiments under harsher conditions. In: 2016 International Conference of the Biometrics Special Interest Group (BIOSIG). pp. 1–6 (Sep 2016), http://ieeexplore.ieee.org/document/7736905/
11. Gunetti, D., Picardi, C.: Keystroke analysis of free text. ACM Transactions on Information and System Security (TISSEC) 8(3), 312–347 (Aug 2005), http://doi.acm.org/10.1145/1085126.1085129
12. Joyce, R., Gupta, G.: Identity authentication based on keystroke latencies. Commun. ACM 33(2), 168–176 (Feb 1990), http://doi.acm.org/10.1145/75577.75582
13. Kang, P., Cho, S.: Keystroke dynamics-based user authentication using long and free text strings from various input devices. Information Sciences 308, 72–93 (2015)
14. Karnan, M., Akila, M., Krishnaraj, N.: Biometric personal authentication using keystroke dynamics: A review. Applied Soft Computing 11(2), 1565–1573 (2011)
15. Killourhy, K.S., Maxion, R.A.: Comparing anomaly-detection algorithms for keystroke dynamics. In: International Conference on Dependable Systems & Networks (DSN-09). pp. 125–134. EEE Computer Society Press, Los Alamitos, California (2009), http://lsia.fi.uba.ar/papers/killourhy09.pdf
16. Leggett, J., Williams, G.: Verifying identity via keystroke characterstics. International Journal of Man-Machine Studies 28(1), 67–76 (1988)
17. Messerman, A., Mustafic, T., Camtepe, S.A., Albayrak, S.: Continuous and non-intrusive identity verification in real-time environments based on free-text keystroke dynamics. In: Biometrics (IJCB), 2011 International Joint Conference on. pp. 1–8. IEEE (2011)
18. Monrose, F., Rubin, A.D.: Authentication via keystroke dynamics. In: Proceedings of the 4th ACM conference on Computer and communications security. pp. 48–56. ACM (1997)
19. Polemi, D.: Biometric techniques: review and evaluation of biometric techniques for identification and authentication, including an appraisal of the areas where they are most applicable. Reported prepared for the European Commision DG XIIIC 4 (1997), http://lsia.fi.uba.ar/papers/polemi97.pdf