Avances en Reconocimiento de Patrones de Tecleo para la Identificación de Personas en Ambientes Web

Jorge Ierache, Hernán Merlino, German Concilio, Enrique Calot, Nahuel González,.

Laboratorio de Sistemas de Información Avanzados, Departamento Computación
Facultad de Ingeniería, Universidad de Buenos Aires.
Av. Paseo Colón 850 - C1063ACV - Buenos Aires -Argentina
Tel +54 (11) 4343-0893 / 4343-0092.

{jierache,hmerlino,ecalot,gconcilio}@lsia.fi.uba.ar

Resumen

El Laboratorio de Sistemas de Información Avanzados (LSIA) de la Facultad de Ingeniería de la Universidad de Buenos Aires (FIUBA) cuenta con la línea de investigación *Dinámica de Tecleo*. En el marco de un Proyecto de desarrollo Estratégico (PDE) se han hecho aportes relevantes al área, donde se destaca la integración de los resultados teóricos obtenidos por el LSIA —en un proyecto UBACYT previo— con la industria. Se demostró el potencial de la aplicación práctica de las técnicas desarrolladas en el contexto de un *e-commerce*, adaptando los algoritmos de autenticación e identificación de personas por su ritmo de tecleo para su utilización en un entorno web. En un entorno web no se tiene un control sobre la población de usuarios que utilizan la plataforma y hay una gran variedad de tipos de dispositivos de entrada que afectan la efectividad final del método.

Palabras clave: *LSIA, Dinámica de tecleo, Keystroke Dynamics, ambientes web, e-commerce*

Contexto

El LSIA fue creado en el 2011 y cuenta con diversas publicaciones [LSIA, 2020]. Dentro de sus líneas de investigación, se le presta especial atención a la línea de *Dinámica de Tecleo*. En 2019 se inició un Proyecto de desarrollo Estratégico (PDE) vinculado a la transferencia a la industria reconocido por la UBA (PDE-44-2019). Su objetivo es estudiar el reconocimiento de patrones de dinámica de tecleo en ambientes web, en particular en el contexto e-commerce.

El proyecto PDE utiliza como base de los algoritmos desarrollados en un proyecto UBACYT 20020130200140BA finalizado. EL proyecto estaba centrado en los métodos de educción de cadencia de tecleo centrado en el contexto emocional de un individuo aplicando Interfaces Cerebro-Máquina (BMI). A partir del mismo, se generaron bases de conocimiento fundamentales en el área de reconocimiento de patrones de tecleo y se sentaron las bases del algoritmo que luego se adaptó al contexto de aplicación práctica propuesto en el proyecto PDE.

Introducción

Los patrones neurofisiológicos que vuelven única a una firma manuscrita se

pueden observar también en el ritmo de tecleo de un usuario [Joyce & Gupta, 1990]. La técnica que analiza este tipo de patrones se llama Dinámica de Tecleo o Keystroke Dynamics [Calot et al. 2014], [González et al. 2015, 2016]. Desde el LSIA se estudian las técnicas de identificación y autenticación de personas por su ritmo de tecleo.

En este artículo se enuncian los avances en materia de reconocimiento de patrones de tecleo para la identificación de personas en ambientes web, en particular en el contexto e-commerce.

Resultados y Objetivos

Se realizó un trabajo colaborativo con una de las principales empresas de *e-commerce* de América Latina, quien dio acceso a un conjunto de datos de experimentación. Se obtuvo un conjunto de datos de más de 2000 usuarios reales la plataforma elegidos al azar, sin sesgo.

La base del algoritmo utilizado para identificar usuarios a partir de su ritmo de tecleo es producto del del proyecto UBACYT 20020130200140BA [Calot et al., 2013, 2014, 2015], [González et al. 2015, 2016]. Se han hecho adaptaciones pertinentes al algoritmo para mejorar su efectividad en los entornos web.

Para aumentar la calidad de los datos de entrenamiento se realizó un refinamiento del lote de pruebas. Se configuró un mínimo de teclas alfanuméricas presionadas en una sesión y se descartaron las sesiones preparadas offline, es decir, cuando la secuencia capturada es sólo una combinación de las teclas CTRL + V. Este tipo de sesiones no reflejan el normal comportamiento del usuario al teclear y en consecuencia las muestras simplemente se descartaron.

Las pausas del individuo al teclear tampoco reflejan su cadencia característica y son consideradas fuentes de ruido. Para reducir esta fuente de ruido de los datos, se particiona el texto de entrada utilizando dos estrategias: por delimitadores de tecla o separador de palabras, y por intervalos de silencio máximos –parámetro configurable en el algoritmo-. Además se filtran todas aquellas particiones con menos de un mínimo de teclas alfanuméricas presionadas.

Las letras y números son generalmente utilizados para escribir secuencias que se repiten con frecuencia y por lo tanto tienden a ser ejecutadas con una cadencia relativamente estable. Contrariamente, la mayoría de las teclas especiales presentan patrones que distan notablemente de la predictibilidad. Sin embargo, la consideración de las teclas especiales en los contextos de las teclas comunes mejora la clasificación.

Para probar el algoritmo de distancia entre usuarios enrolados se asumió que la persona real detrás de un usuario es siempre es la misma. Para cada usuario, sus sesiones se subdividieron en dos subconjuntos mutuamente exclusivos, etiquetados como A y B. De esta manera, generamos dos patrones con la misma identidad de usuario.

Utilizamos varias muestras de usuarios para evaluar el comportamiento de distintos métodos de cálculo de distancia entre usuarios con distintos aspectos personalizables. Por ejemplo, para la un grupo A inicial de 40 usuarios, se obtuvieron 80 patrones, dos por usuario, uno perteneciente al grupo ~A y otro a ~B. A través del modelado de cadencia de tecleo de cada patrón, intentamos descubrir qué otro patrón es más probable que comparta el mismo identificador de usuario. Contamos con un éxito cuando identificamos el usuario del conjunto ~A con la menor distancia de patrón para el mismo usuario del conjunto B. Además, se estudió cómo influyen la cantidad de texto

ingresado por el usuario y la cantidad de usuarios enrolados en la plataforma en el porcentaje de acierto del método.

Se ha logrado demostrar que se puede distinguir a una persona de un lote de 2000 usuarios con una efectividad superior al 89% utilizando únicamente nuestro algoritmo de identificación por cadencia de tecleo [Concilio et al., 2018]. Además, en un conjunto de 100 usuarios seleccionados al azar se pudo autenticar sesiones con un EER de 7,82 %. El resultado obtenido es comparable con resultados de otros autores del estado del arte, a pesar de que el set de datos presenta dificultades extra, como la no heterogeneidad de los teclados usados para tipiar, sistemas operativos o el nivel educativo de los individuos.

Por último, se propuso un algoritmo novedoso para asociar usuarios con una huella biométrica característica de la persona real. Para probar este algoritmo se generó un modelo de cadencia de tecleo para cada usuario del conjunto de pruebas. Luego se calculó la distancia definida por el algoritmo a cada usuario del conjunto de datos y se tomó un umbral de sensibilidad debajo del cual dos usuarios se dice que la misma huella biométrica. Decimos que un usuario tiene un cruce con otro usuario cuando comparten la misma huella biométrica.

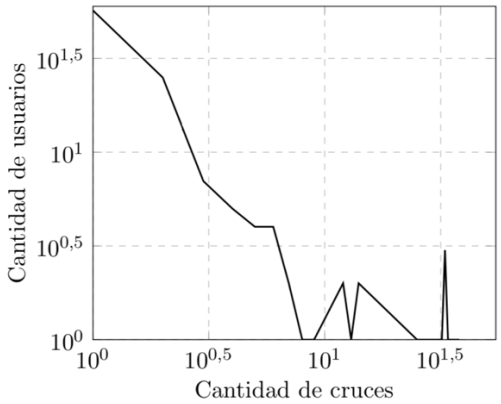

***Figura 1**. Distribución de la cantidad de cruces generados por el algoritmo*

La Figura 1 muestra que aproximadamente un 60% de los usuarios no tienen cruces con ningún usuario –notar que siempre un usuario cruza consigo mismo por definición del experimento–. En un proceso de revisión empírica manual se determinó que este grupo de usuarios en su mayoría no tienen indicios de compartir cuenta con otro usuario de la plataforma.

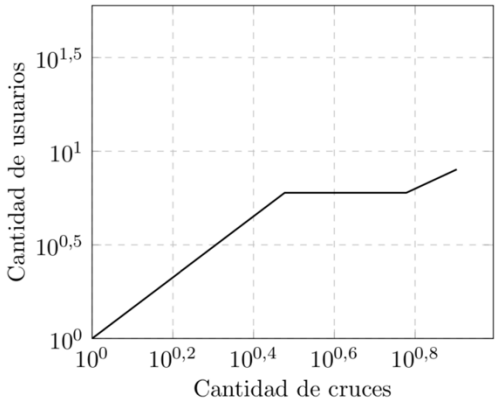

***Figura 2**. Distribución de cantidad de cruces para los usuarios con comportamiento fraudulento previamente detectado.*

Los usuarios con más de 7 cruces representan el 8% de la población de usuarios evaluada. Se determinó en un proceso de revisión empírica que existía realmente una misma persona física detrás de esos usuarios. Dentro de este grupo de usuarios con muchos cruces, se lograron aislar 30 usuarios con comportamiento fraudulento confirmados. Se concluye que los resultados experimentales muestran que es posible detectar usuarios controlados por una misma persona física, con interés fraudulento, a través de patrones de tecleo compartidos por varios usuarios de la plataforma [Concilio et al., 2018].

En otro estudio se ha logrado demostrar que la distancia euclediana es la mejor opción a la hora de querer mitigar el ruido producido por las emociones de las personas al momento de teclear [Calot, et al., 201]. Además, en [Calot et al., 2019 b]

de han evaluado mejoras al algoritmo base original, concluyendo que una distancia específica de Minkowsky reduce el EER de 21,9% a 17,4% con el set de datos de experimentación en condiciones reales.

Formación de Recursos Humanos

El laboratorio actualmente se conforma de dos investigadores formados, un investigador formado invitado, dos estudiantes de doctorado, un alumno investigador. **S**e han radicado en la temática de cadencia de tecleo una tesis de doctorado y una tesis de grado de la Facultad de Ingeniería de la UBA ambas finalizadas y defendidas.

Referencias

CALOT, E. 2015. "Keystroke Dynamics keypress latency dataset". Base de datos para investigación. http://lsia.fi.uba.ar/pub/papers/kd-dataset/

CALOT, E.; PIRRA, F.; RODRIGUEZ, J.M.; PEREIRA, G.; IRIBARREN, J.; IERACHE, J. 2014. "Métodos Adaptativos de Educción de Dinámica de Tecleo Centrado en el Contexto Emocional de un Individuo aplicando Interfaz Cerebro Computadora". XVI Workshop de Investigadores en Ciencias de la Computación, ISBN 978-950-34-1084-4.

CALOT, E.; RODRIGUEZ, J.M.; IERACHE, J. Improving versatility in keystroke dynamic systems. En Proceedings del XIX Congreso Argentino de Ciencias de la Computación, number 5606, 2013. ISBN 978-987-23963-1-2. URL http://hdl.handle.net/10915/32428

CALOT, E.; RODRIGUEZ, J.M.; IERACHE, J., 2014. Improving versatility in keystroke dynamic systems. En Jorge Raúl Finochietto y Patricia Mabel Pesado, editors, Computer Science & Technology Series. XIX Argentine Congress of Computer Science, Selected papers, páginas 289–298. Editorial de la Universidad Nacional de La Plata (EDULP), 2014b. ISBN 978-987-1985-49-4. URL http://lsia.fi.uba.ar/papers/calot14b.pdf

CALOT, E., ROSSI F., GONZÁLEZ N., HASPERUÉ, W., IERACHE, J.. Avances en educción de dinámica de tecleo y el contexto emocional de un individuo aplicando interfaz cerebro computadora. En WICC 2016, Entre Ríos, Argentina), páginas 872–876, jun. 2016. ISBN 978-950-698-377-2. URLhttp://hdl.handle.net/10915/53247

CALOT, E., IERACHE, J.. Multimodal biometric recording architecture for the exploitation of applications in the context of affective computing. En Proceedings del XXIII Congreso Argentino de Ciencias de la Computación (La Plata, 2017), number 10529, páginas 1030–1039, 2017. ISBN 978-950-34-1539-9. URL http://hdl.handle.net/10915/63866

CALOT, E., IERACHE, J., HASPERUÉ, W.. 2019. Robustness of keystroke dynamics identification algorithms against brain-wave variations associated with emotional variations. En Advances in Intelligent Systems and Computing. Springer, c. En prensa

CALOT, E., IERACHE, J., HASPERUÉ, W.. 2019. Document Typist Identification by Classification Metrics Applying Keystroke Dynamics Under Unidealised Conditions. En International Conference on Document Analysis and Recognition Workshops (ICDARW) IEEE, 2019. p. 19-24.

CONCILIO, G., IERACHE, J., MERLINO, H., CALOT, E.. Application of Keystroke Dynamics Modelling Techniques to Strengthen the User Identification in the Context of E-commerce. En XXIV CACIC 2018. En prensa. URL http://lsia.fi.uba.ar/papers/concilio18.pdf

JOYCE, R.; GUPTA, G. 1990. Identity authentication based on keystroke latencies. Commun. ACM 33, 2 (February 1990), 168-176. http://doi.acm.org/10.1145/75577.75582

GONZÁLEZ, N., CALOT , E. Of Keystroke Dynamics In Free Text. En Biometrics Special Interest Group (Biosig), 2015 International Conference Of The, Páginas 1–5, Sep. 2015. ISBN 978-3-88579-639-8. DOI: 10.1109/Biosig.2015.7314606

GONZÁLEZ, N., CALOT, E. Y IERACHE, J... A Replication Of Two Free Text Keystroke

Dynamics Experiments Under Harsher Conditions. En 2016 International Conference Of The Biometrics Special Interest Group (Biosig), Páginas 1–6, Sep. 2016. DOIi: 10.1109/Biosig.2016.7736905

LSIA: http://lsia.fi.uba.ar vigente marzo 2020