LA INVESTIGACIÓ EN L'ÀMBIT EDUCATIU - DEBATS DE RECERCA 15: 115-126 (2023)
Societat Andorrana de Ciències, Andorra
DOI: 10.2436/16.0110.21.121. ISBN: 978-99920-61-70-1

Aleix Dorca i Josa

Doctor en Ciències de la Informació per la Universitat d'Andorra, llicenciat en Informàtica, màster en Seguretat Informàtica i Programari Lliure, i responsable dels serveis informàtics de l'UdA

Aleix Dorca i Josa

La dinàmica de tecleig com a eina per identificar usuaris en entorns educatius

L'ús de teclats tradicionals i la possibilitat de mesurar el ritme particular que té un usuari a l'hora d'escriure en un aplicatiu informàtic permeten la possibilitat d'identificar aquests usuaris. Al llarg dels anys, l'obtenció de mètodes d'autenticació, identificació i verificació robustos ha estat el focus principal en el camp de la dinàmica de tecleig.

Aquest article vol mostrar una alternativa als mètodes tradicionals per determinar fins a quin punt es pot establir la identitat dels usuaris quan utilitzen recursos en línia com ara entorns d'aprenentatge en línia. Aquesta investigació s'ha dut a terme en un entorn real utilitzant una metodologia de text lliure. El mètode proposat se centra en la hipòtesi que la posició d'una combinació particular de lletres en una paraula és d'alta importància. El model de l'usuari es construeix utilitzant els intervals entre les successives pulsacions de tecles i el context de les paraules escrites, és a dir, tenint en compte on ha tingut lloc un traç de lletres determinat.

Els resultats haurien d'ajudar a determinar si l'ús de la dinàmica de tecleig i el mètode proposat és suficient per identificar els usuaris a partir del contingut que creen amb un nivell de certesa prou correcte. Aleshores, es podria utilitzar com a mètode per garantir que un usuari no sigui suplantat per un altre, en esquemes d'autenticació, o per ajudar a determinar l'autoria de diferents parts d'un document escrit per més d'un usuari.

1. Introducció

Identificar correctament els usuaris és un dels principals objectius quan s'utilitzen tècniques biomètriques. En entorns d'aprenentatge en línia pot aparèixer el dubte de si un treball ha estat escrit per l'usuari que l'ha enviat. També, durant la realització d'exàmens en línia, sobretot en aquests darrers anys de confinament, la seguretat que l'usuari és qui diu ser és necessària. Tenir, únicament, els usuaris autenticats a la

plataforma en línia o fins i tot al seu entorn d'escriptori no és cap garantia que els treballs presentats hagin estat escrits per aquests usuaris.
La dinàmica de tecleig s'estudia des de finals dels anys setanta i pot ser d'ajuda en el problema plantejat. Aquest camp d'estudi s'ha dividit en diferents branques, i l'autenticació, la identificació i la verificació són les més rellevants. Al mateix temps, l'estudi de com els usuaris escriuen al teclat s'ha dut a terme utilitzant dos enfocaments principals: text fix i text lliure. Un exemple típic de text fix seria el d'una contrasenya, quelcom conegut pels usuaris que sempre escriuen de la mateixa manera. S'oposa a la idea de la metodologia de text lliure en la qual els usuaris poden escriure sense restriccions d'extensió o contingut.
La metodologia tradicional consisteix a crear un model de les característiques que millor descriuen un usuari quan tecleja en un teclat. Contra aquest model es poden comparar noves mostres per verificar-ne la validesa, sempre amb un cert nivell d'error. La normativa europea per a sistemes de control d'accés especifica una taxa de falsa alarma inferior a l'1%, amb una taxa de fallada no superior al 0,001%.
Aquest estudi utilitza les dades de context de les paraules entrades per identificar els usuaris en contraposició a altres tècniques conegudes com, per exemple, la freqüència n-graf. Aquest mètode discuteix si el ritme d'un usuari en particular és el mateix quan escriu, per exemple: ES, MES, TESI o MESOS. La combinació de lletres E-S normalment es consideraria un dígraf i s'agruparia en una estructura de dades sense tenir en compte si ha aparegut al principi, al mig o al final de la paraula, o la mida d'aquesta paraula. Aquestes particularitats no s'havien estudiat abans tot i que s'havien proposat com una possible línia de treball.

2.Estat de l'art

El camp de recerca basat en text lliure ha estat molt menys estudiat que l'alternativa de text fix. En un dels primers articles que tractaven text lliure, els resultats no van ser gaire prometedors amb només un 23% d'identificació positiva. L'àmplia gamma d'entorns diferents d'entrada de dades (sovint molt adaptats i controlats), nombre d'usuaris i mostres, mètodes de classificació i altres factors fa que sigui molt difícil establir un estàndard on comparar i més encara quan s'estudien

característiques menys estudiades, com les dades de context.
Un dels articles més citats és el de D. Gunetti i C. Picardi . Els autors van calcular distàncies relatives i absolutes entre mostres i les van combinar per obtenir els resultats. Van obtenir un False Acceptance Rate (FAR) del 0,005% i un False Rejection Rate (FRR) del 5%. Un dels problemes del seu mètode era que els recursos necessaris per obtenir el grau de desordre d'un vector podien ser exigents. Altres estudis han intentat tractar aquest problema d'escalabilitat.
La influència de diferents teclats és quelcom que també s'ha estudiat. L'estudi d'M. Villani *et al*. dut a terme és d'alta rellevància per tal d'avaluar els resultats presentats en aquest estudi. La identificació de l'usuari era més precisa si l'usuari utilitzava sempre el mateix teclat (taxa d'identificació del 99,8%). En aquest estudi, els usuaris no només podien enviar informació des de qualsevol dispositiu, sinó també des de qualsevol ubicació, de manera que els resultats s'han pogut veure afectats per aquesta manca de consistència.
Un altre estudi critica la metodologia basada en n-grafs suggerint que aquesta estructura de dades no proporciona prou informació sobre la forma en què un usuari escriu. L'estudi suggereix que les paraules senceres podrien donar resultats iguals o millors que l'ús de n-gràfics curts. El present estudi respondrà a la pregunta de si la longitud importa mitjançant l'anàlisi de diferents longituds de paraules.
La metodologia utilitzada en aquest treball comparteix similituds amb el treball de Messerman *et al*. i amb M. Curtin *et al*. Tots dos van utilitzar mostres de n-grafs variables per construir els models. Un resultat interessant de la seva recerca va ser el fet que si es comparava una nova mostra amb un nombre creixent de models, la possibilitat d'identificar correctament l'usuari disminuïa ràpidament.
Brizan *et al*. van publicar un article en què van intentar identificar la demografia dels usuaris estudiats amb una precisió del 82.2% quan les mostres tenien almenys 50 paraules. Això va en consonància amb el que s'ha trobat en la investigació presentada en aquest article.
Val la pena assenyalar que aquesta tècnica biomètrica s'ha utilitzat en esquemes multimodals que inclouen, entre altres coses, el reconeixement facial i el reconeixement de veu per millorar la identificació global.
També és important destacar que durant la pandèmia de la

COVID-19 no ha estat estrany, per part de les universitats, el fet d'haver d'adaptar-se a entorns en línia per fer exàmens o treballs. La Universitat Oberta de Catalunya,[1] per exemple, ha fet recerca en aquest àmbit, sobretot en el reconeixement facial, però també en dinàmica de tecleig.[2]

3. Metodologia

3.1. Recol·lecta de mostres

Les mostres es van recollir durant diferents períodes semestrals a partir dels missatges enviats als fòrums del Campus Virtual de la Universitat d'Andorra. Cadascun d'aquests missatges l'anomenem *sessió* (S). Els intervals de temps per a cada tecla premuda es van recollir i es van enviar de manera segura a un servidor remot on es van emmagatzemar en una base de dades. Per a cada esdeveniment la informació recopilada era: un usuari i un identificador de sessió, el codi tecla de l'esdeveniment, el tipus d'esdeveniment (Keydown o Keyup), la marca de temps i altres metadades sobre el dispositiu i ubicació de l'usuari. L'usuari va estar informat en tot moment del tractament de dades d'aquest estudi.

Es van analitzar un total de 60 usuaris, van ser seleccionats entre els que havien enviat un nombre d'esdeveniments més gran al Campus Virtual. Es van avaluar prop de 4.000 sessions. Val la pena assenyalar que la informació només es va recopilar d'ordinadors d'escriptori.

El perfil dels usuaris seleccionats va ser força heterogeni, una característica molt apreciada en aquest tipus d'estudis. Es van recollir mostres d'estudiants i professorat de tota mena d'estudis que s'ofereixen a la universitat. Finalment, l'edat dels usuaris estava compresa entre els 18 i els 65 anys.

3.2. Anàlisi d'intervals

L'estudi d'una sessió (S) consisteix a analitzar els diferents esdeveniments Keydown (KD) i Keyup (KU) per trobar els intervals de temps entre ells. Això permet obtenir la informació dels intervals Press–Release (també coneguts com *dwell time* o PR) i Release–Press (també coneguts com *fly time* o RP). El procés de detecció de paraules es va fer tenint en compte dues característiques bàsiques: delimitadors coneguts (és a dir: espai, coma, punt…) i un interval de temps màxim de silenci (300ms) obtingut empíricament.

1-Universitat Oberta de Catalunya: https://www.uoc.edu
2- TeSLA: https://tesla-project-eu.azurewebsites.net

La figura [fig:intervals] mostra un exemple dels intervals de temps de les paraules: THE SUN. La primera paraula (THE) està formada pels intervals PR següents: D_1=54, D_2=28 i D_3=18. Els intervals RP són: F_1=25 i F_2=5. Quan es detecta un separador de paraules (un espai, en aquest cas) es descarten els intervals d'aquest esdeveniment (F_3, D_4 i F_4). La segona paraula (SUN) està formada pels següents intervals PR: D_5=32, D_6=38 i D_7=28 i dels següents intervals RP: F_5=29 i F_6=33. A partir d'aquesta informació també es poden obtenir fàcilment altres intervals com ara Press–Press (PP) o Release–Release (RR).

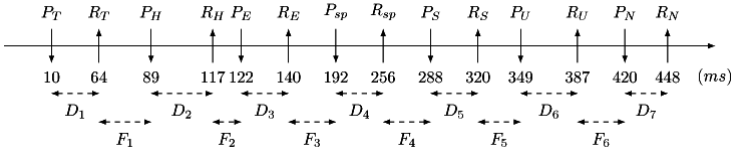

Figura 1: Intervals de temps per a les paraules: THE SUN

3.3. El model d'arbre

Les paraules detectades s'emmagatzemen en una estructura d'arbre com la que es mostra a la figura [fig:tree]. En aquest exemple, s'han afegit a l'arbre les paraules següents: A, T, W, ALL, ALBERT, THE, THERE, THIS, WORD i WIT. Cadascun dels nodes que contenen una lletra poden tenir intervals de tipus PR i RP (primera i segona llista, respectivament). Atès que aquesta investigació va utilitzar quatre característiques (és a dir, PR, RP, PP i RR) es van descartar paraules d'una lletra. Aquestes últimes semblaven afegir poca informació valuosa.
En el model d'arbre, un node pot tenir intervals PR i RP o no, depenent de si l'usuari ha escrit alguna vegada aquesta paraula sencera en particular. La informació de temps s'emmagatzema sempre en el node corresponent a la darrera lletra de la paraula. Si es detecta una paraula més d'una vegada, hi haurà una llista de PR i RP diferent per a cada instància trobada (per exemple: ALL a la figura). Si es detecta una sub-paraula d'una paraula ja emmagatzemada, hi haurà informació de temporització PR i RP en un node que no sigui fulla (per exemple: THE - THERE a la figura [fig:tree]).

Val la pena assenyalar que el model d'arbre es va netejar d'instàncies de paraules fora de tres desviacions estàndard per evitar soroll excessiu.

3.4. Avaluació de sessions

Un cop construït el model d'arbre es van poder comparar noves sessions contra ell i mirar d'establir l'autor d'una sessió determinada. El procés consisteix a cercar cada paraula d'una sessió nova en el model d'arbre i calcular la distància entre la paraula d'origen i la paraula trobada en el model.

Per a aquest estudi es va utilitzar la mesura de distància *Txebixev*. Per obtenir la distància entre una paraula i un model es necessitava un vector origen i un vector objectiu. El vector origen era la llista de temps d'interval a partir de la paraula que es buscava i el vector objectiu era el que s'obtenia de la informació emmagatzemada en el model d'arbre. Si una paraula del model d'arbre tenia més d'una instància s'utilitzava el vector amb la mitjana de totes les instàncies registrades.

En cercar paraules en el model d'arbre es troba una de les situacions següents:

- La paraula no es troba en el model. Es descarta.
- La paraula s'ha trobat en el model completament i la darrera lletra ha estat la d'un node fulla. La distància es pot obtenir de manera immediata.

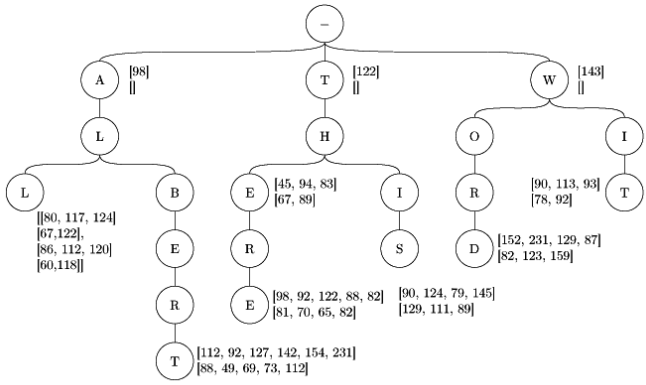

Figura 2: Model d'arbre

• La paraula s'ha trobat parcialment però el node en què s'ha trobat l'última lletra de la paraula origen no conté informació d'intervals de temps perquè és la primera vegada que l'usuari escriu aquesta paraula sencera en particular. Els temps parcials de les fulles del node de l'última lletra es poden determinar i utilitzar-los per trobar la distància entre aquestes sub-paraules parcials.

• La paraula s'ha trobat parcialment en el model però encara quedaven lletres de la paraula origen per trobar. Anteriorment, l'usuari només havia introduït paraules més curtes amb les mateixes lletres arrel. En aquest cas, només s'utilitzen els temps de la sub-paraula parcial trobada.

3.5. Paràmetres estudiats en relació amb el context

S'han estudiat els següents paràmetres per veure el seu efecte a l'hora d'avaluar les dades de context:

• Longitud de les paraules: aquest paràmetre analitza si totes les longituds de paraules del model d'arbre són igual de rellevants. Això és d'interès, no només en termes de rendiment i optimització de models, sinó també per determinar si els usuaris tenen una tendència natural a ser més consistents en la seva escriptura per a un nombre limitat de pulsacions de tecles. S'han provat els següents valors: nombre il·limitat de lletres (≥2); superior a 2 (>2); entre 2 i 5 ([2–5]); i entre 3 i 7 ([3–7]).

• Nombre de paraules que es troben en cercar el model. Un problema recurrent va aparèixer quan el nombre de paraules d'una sessió era massa baix. Podria passar que un usuari només hagués accedit al fòrum per contribuir amb unes paraules. Aquest paràmetre intenta mitigar aquest problema establint un nombre mínim de paraules ja sigui en la sessió que s'analitza o en el model. S'ha establert un valor llindar de 50 paraules.

3.6. Identificant l'usuari d'una sessió

La distància Txebixev entre dos vectors $\vec{X}$ i $\vec{Y}$ es defineix per la següent equació:

$$DCh(\vec{X},\vec{Y})=max_{i=1}^{n}|X_i-Y_i|$$

Cada sessió S té W paraules. Cada paraula W_i és un vector de valors $\vec{X}$. Aquest vector pot incloure una combinació de dwell times i/o fly times dels intervals de temps registrats en funció de la característica F que s'analitzi. F pot ser un dels següents:

PR (Press–Release), RP (Release–Press), PP (Press–Press), i RR (Release–Release).

La paraula W_i cercada en el model M pertanyent a l'usuari U produeix un altre vector $\vec{Y}$.A partir d'aquests dos vectors $\vec{X}$ i $\vec{Y}$ es pot determinar la distància. A partir d'aquestes distàncies es calculen dos valors: la mitjana md i la mitjana ponderada wmd per a totes les característiques. Els valors md i wmd fan ús de la profunditat d en què es troba cada W_i. La mitjana ponderada s'obté mitjançant els pesos següents: tots els valors fins a 100 tenen un pes de 15; els valors entre 100 i 200 tenen un pes de 5; i els valors entre 200 i 500 tenen un pes d'1. Es descarten valors superiors a 500. Aquests pesos es van obtenir empíricament.

En aquest punt, es disposa d'una $md(W_i)$ i un valor $wmd(W_i)$ per a cada paraula W_i cercada en el model M. La distància global final gd entre una sessió S i el model M es compon de quatre valors (gdm, $gdmed$, $gdwm$, $gdwmed$) calculats mitjançant el mètode següent:

$$\forall md(W_i)\in S, gd_m = Mean(md(W_i)), gd_{med} = Median(md(W_i))$$
$$\forall wmd(W_i)\in S, gd_{wm} = Mean(wmd(W_i)), gd_{wmed} = Median(wmd(W_i))$$

Com a exemple del mètode proposat, la taula 1 mostra els resultats després d'haver calculat la mesura de la distància Txebixev entre les paraules d'una sessió d'origen i el model d'arbre d'un usuari. En aquest exemple es mostren tres usuaris diferents (columna Test). Cada usuari ha tingut quatre paraules comparades (here, sun, there, i moon) entre la sessió d'origen i el model d'arbre. Es mostren els valors de distància de les quatre característiques utilitzades (PP, RP, PP, i RR). La columna Real identifica el propietari real de la sessió. La columna Profunditat mostra el nombre de lletres que s'han trobat en el model d'arbre. Si la paraula origen només s'hagués trobat parcialment aquest valor mostraria la profunditat a la qual s'havia trobat l'última lletra. Finalment, les columnes md i wmd mostren els valors mitjana i mitjana ponderada calculats per a cada paraula.

Per a la primera fila de l'usuari 3207 el valor mitjà seria: (69+144+176+99)/4=122. De la mateixa manera, el valor mitjà ponderat seria: (69·15+144·5+176·5+99·15)/40=103. Aquests dos valors es divideixen aleshores per la profunditat a la qual s'ha trobat l'última lletra de la paraula: md=122/4=30.50 i wmd=103/4=25.75.

Paraula	Característica				Profunditat	Usuari		md	wmd
	PR	RP	PP	RR		Test	Real		
here	69	144	176	99	4	3207	192	30.50	25.75
sun	67	19	48	21	3	3207	192	12.92	12.92
there	56	135	145	93	5	3207	192	21.45	18.18
moon	88	33	66	30	4	3207	192	13.56	13.56
here	84	200	163	124	4	37	192	35.69	30.79
sun	71	16	58	74	3	37	192	18.25	18.25
there	72	187	145	110	5	37	192	25.70	21.93
moon	66	25	70	60	4	37	192	13.81	13.81
here	23	11	16	20	4	192	192	4.38	4.38
sun	15	15	14	23	3	192	192	5.58	5.58
there	34	20	13	18	5	192	192	4.25	4.25
moon	20	30	15	28	4	192	192	5.81	5.81

Taula 1. Distàncies després de comparar una sessió contra tres models de diferents usuaris

Finalment, a partir de cadascun d'aquests valors md i wmd i per a cada usuari U es calculen els quatre valors finals gdm, gd_{med}, gd_{wm}, gd_{wmed}. La taula 2 mostra aquests valors finals per a l'exemple proposat. Per exemple, per a l'usuari 3207, gdm=(30.50+12.92+21.45+13.56)/4=19.61

Usuari		gd_m	gd_{med}	gd_{wm}	gd_{wmed}	Vots
Test	Real					
3207	192	19.61	17.60	17.51	15.87	0
37	192	23.36	21.20	21.98	20.09	0
192	192	5.01	5.01	4.98	4.98	4

Taula 2. Valors finals per al mètode proposat

3.7. Fusió utilitzant un mètode de votació

A la taula 2, la columna Vots mostra el nombre total on cadascun dels valors gd era un mínim en comparació amb els altres usuaris. S'ha observat que en avaluar les sessions utilitzant aquests quatre valors de gd, hi havia algunes sessions identificades incorrectament, però la majoria de les vegades aquests errors no serien reportats pels quatre valors de gd a la vegada. S'ha decidit utilitzar un mètode de fusió per intentar millorar la taxa global d'identificació mitjançant l'ús d'un esquema de votació. Una sessió es determina com a propietat d'un usuari seleccionant si té la majoria dels valors mínims de gd. A l'exemple de la taula 2, l'usuari 192 obté els vots dels 4 mètodes i, per tant, es determina com a propietari de la

sessió. Tenir un nombre parell de característiques no ajuda sempre a discriminar totalment. Queda com a treball futur millorar aquest esquema.

3.8. Generalització

Per provar el mètode proposat es van utilitzar trenta conjunts de proves escollits aleatòriament de 40 usuaris del grup disponible de 60 usuaris. La partició de sessions per provar i construir els models va ser del 30/70% Cada sessió es va comparar amb tots els models. Aquest procés es va repetir per a cada sessió de cada usuari. A continuació, es va determinar el percentatge de sessions correctament identificades. Per obtenir el millor resultat, també es mostren els valors mitjans FAR i FRR.

4. Resultats

Els resultats mostren l'efecte dels paràmetres analitzats relacionats amb el context (longitud de la paraula i recompte mínim de paraules trobat per sessió). La taula 3 mostra també el valor mitjà del percentatge de sessions correctament identificades quan es van utilitzar cadascun dels valors de gd i el sistema de votació.

Taula 3. Resultats per característiques i mètodes

Paraules	Sense límit inferior				>50			
Mida	≥2	>2	[2−5]	[3−7]	≥2	>2	[2−5]	[3−7]
Mètode								
gd_m	83.98	83.43	80.26	81.59	97.48	96.60	96.00	95.57
gd_{med}	84.78	83.73	81.31	82.07	98.28	96.93	97.18	96.22
gd_{ucm}	86.14	86.17	83.23	84.87	98.10	98.10	97.47	97.42
gd_{ucmed}	85.02	84.68	81.87	83.12	98.18	97.71	97.41	96.97
Votació	87.90	87.29	85.22	86.08	98.74	98.44	98.18	97.95

El millor valor de la taula 3 és un percentatge de 98.74% correctament identificat. Amb un valor mitjà de 377 sessions respecte als models, el FRR va ser de 0,0126 i el FAR de 0,0002.

Aquest resultat es va obtenir utilitzant totes les longituds de les paraules. Al llarg de la taula, es pot veure que descartar longituds de paraules més grans no millora els resultats. D'altra banda, si l'optimització i el rendiment de l'ordinador són una preocupació important, la diferència en el nombre de

Referències

CENELEC (2002). *European Standard EN 50133-1: Alarm systems. Access control systems for use in security applications. Part 1: System requirements.*

BOURS, P. (2012). Continuous keystroke dynamics: A different perspective towards biometric evaluation. *Information Security Technical Report* 17.1, 36 - 43.

SIM, T., ZHANG, S., JANAKIRAMAN, R., & KUMAR, S. (2007). Continuous verification using multimodal biometrics. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29.4, 687 - 700.

MONROSE, F., & RUBIN, A. D. (1997). Authentication via keystroke dynamics. *Proceedings of the 4th ACM conference on Computer and communications security*. ACM, 48 - 56.

ALSULTAN, A., & WARWICK, K. (2013). Keystroke Dynamics Authentication: A Survey of Free-text Methods. *International Journal of Computer Science Issues* 10.4.

GUNETTI, D., & PICARDI, C. (2005). Keystroke Analysis of Free Text. *ACM Transactions on Information and System Security* 8.3, 312 - 347. ISSN: 1094-9224.

HU, J., GINGRICH, D., & SENTOSA, A. (2008). A k-nearest neighbor approach for user authentication through biometric keystroke dynamics. *Communications, 2008. ICC'08. IEEE International Conference on.* IEEE, 1556 - 1560.

VILLANI, M. *et al.* (2006). Keystroke biometric recognition studies on long-text input under ideal and application-oriented conditions. *Computer Vision and Pattern Recognition Workshop, 2006. CVPRW'06. Conference on*. IEEE, 39 - 39.

sessions correctament identificades quan s'utilitzen totes les longituds de les paraules i quan només s'utilitza l'interval [2–5], per exemple, és marginal.

Sense cap dubte, el paràmetre més important és el nombre mínim de paraules que es troben al model. Quan aquest model s'estableix a un mínim de 50 paraules els resultats milloren substancialment. El desavantatge d'establir aquest paràmetre és que s'avaluen un nombre més baix de sessions.

5. Conclusions

L'objectiu d'aquest estudi era esbrinar si l'ús de la dinàmica de tecleig i les dades de context, a diferència d'altres tècniques conegudes, era un mètode eficaç a l'hora d'intentar identificar els usuaris en entorns educatius, com pot ser el Campus Virtual de la Universitat d'Andorra. S'ha proposat una nova estructura de dades, basada en arbres de paraules, lletres i vectors de distàncies. Dels resultats obtinguts se'n poden derivar les següents conclusions:

- El resultat més important és la validesa de les dades de context com a característica d'identificació. S'ha demostrat que, utilitzant un entorn altament hostil, l'ús de tècniques estadístiques senzilles ofereix una molt bona taxa de precisió, comparable, si no millor, a estudis previs en entorns similars.
- El millor resultat de la longitud de les paraules va ser utilitzar totes les longituds de paraules disponibles.
- Quan hi ha un nombre mínim de paraules que es troben al model, en lloc d'acceptar qualsevol sessió de mida per comparar-la amb els models, els resultats són molt millors. Això concorda amb el que també han manifestat altres estudis.
- El mètode de fusió basat en l'esquema de votació proposat sempre millora els resultats en comparació amb els valors parcials de gd. A partir d'aquestes, la mitjana ponderada i la mediana solen ser les que funcionen millor.

6. Treball futur

A continuació es mostren algunes idees per continuar amb la línia de recerca iniciada en aquest estudi:

- Estudiar si altres característiques de l'usuari com ara l'edat, el sexe, etc., són rellevants a l'hora d'identificar els usuaris. Com que hi ha usuaris de tot tipus d'edats disponibles i també hi ha altres metadades disponibles, es podria provar la segmentació.

Sim, T., & Janakiraman, R. (2007). Are digraphs good for free-text keystroke dynamics?. *Computer Vision and Pattern Recognition, CVPR'07. IEEE Conference on*. IEEE. 2007, 1 - 6.

Messerman, A., Mustafic, T., Camtepe, S. A., & Albayrak, S. (2011). Continuous and non-intrusive identity verification in real-time environments based on free-text keystroke dynamics. *Biometrics (IJCB), 2011 International Joint Conference on*. IEEE, 1 - 8.

Curtin, M. *et al.* (2006). Keystroke biometric recognition on long-text input: A feasibility study. *Proc. Int. MultiConf. Engineers & Computer Scientists (IMECS)*.

Brizan, D. G. *et al.* (2015). Utilizing linguistically-enhanced keystroke dynamics to predict typist cognition and demographics. *International Journal of Human-Computer Studies*.

Giot, R., Hemery, B., & Rosenberger, C. (2010). Low cost and usable multimodal biometric system based on keystroke dynamics and 2d face recognition. *Pattern Recognition (ICPR), 2010 20th International Conference on*. IEEE, 1128 - 1131.

Montalvao, F., Jugurta, R., & Freire, E. O. (2006). Multimodal biometric fusion-joint typist (keystroke) and speaker verification. *Telecommunications symposium, 2006 international*. IEEE, 609 - 614.

Baró-Solé, X. *et al.* (2018). Integration of an adaptive trust-based e-assessment system into virtual learning environments—The TeSLA project experience. *Internet Technology Letters* 1.4, e56.

Dorca Josa, A., Santamaría Pérez, E., & Morán Moreno, J. A. (2016). Identificación de usuarios mediante dinámica de tecleo en entornos de entrada libre usando información de contexto. *XXXI Simposium Nacional de la Unión Científica Internacional de Radio (URSI, 2016).*

• Cercar altres funcions, mètodes i estratègies per augmentar el percentatge de sessions correctament identificades sense haver de sacrificar sessions pobres o més curtes.
• Per millorar el rendiment del sistema, i veient que en la majoria dels casos triar un paràmetre sobre un altre dóna poca millora en els resultats, es podrien establir algunes restriccions a l'hora de construir el model d'arbre.
• Es podria analitzar si els paràmetres estudiats són vàlids per a tots els usuaris de la mateixa manera o si grups d'usuaris són més susceptibles a certs paràmetres.