Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 30ου Πανελληνίου Συνεδρίου Στατιστικής (2017), σελ.80-95

ΕΛΕΓΧΟΣ ΗΜΕΡΗΣΙΑΣ ΧΡΗΣΗΣ ΤΟΥ ΥΠΟΛΟΓΙΣΤΗ ΔΙΑ ΜΕΣΩ ΔΥΝΑΜΙΚΗΣ ΤΗΣ ΠΛΗΚΤΡΟΛΟΓΗΣΗΣ

Τσιμπερίδης Ιωάννης, Καράκος Αλέξανδρος
Δημοκρίτειο Πανεπιστήμιο Θράκης
itsimper@ee.duth.gr, karakos@ee.duth.gr

ΠΕΡΙΛΗΨΗ

Το Διαδίκτυο αποτελεί ένα εργαλείο, με το οποίο σήμερα συνδέονται δισεκατομμύρια άνθρωποι με σκοπό την επικοινωνία, τη διασκέδαση, την εργασία και τη μόρφωση. Παράλληλα όμως με τα οφέλη που αποκομίζονται, ελλοχεύουν και κίνδυνοι που μπορεί να βλάψουν τους χρήστες οικονομικά, ηθικά και κοινωνικά. Πολλοί από αυτούς τους κινδύνους προέρχονται από κακόβουλους χρήστες, των οποίων ένα από τα σημαντικότερα όπλα είναι η ανωνυμία που μπορούν να διατηρήσουν στο χώρο του Διαδικτύου. Στην παρούσα εργασία προτείνεται μία μέθοδος, η οποία αξιοποιεί τα δεδομένα που παράγονται από τον τρόπο που πληκτρολογούν οι χρήστες. Σκοπός είναι η εύρεση κάποιων χαρακτηριστικών τους, έτσι ώστε να αρθεί η πλήρης ανωνυμία και να είναι εφικτή η προειδοποίηση των ανυποψίαστων χρηστών για την πιθανότητα ο συνομιλητής τους να έχει παραποιήσει τα χαρακτηριστικά του. Το προτεινόμενο σύστημα υλοποιήθηκε με τεχνητά νευρωνικά δίκτυα και το ποσοστό ορθής πρόβλεψης ανήλθε στο 75%, υπερβαίνοντας κατά πολύ αυτό της τυχαίας πρόβλεψης.

Λέξεις Κλειδιά: Δυναμική της Πληκτρολόγησης, Κατηγοριοποίηση Χρηστών, Μάθηση Μηχανής, Εγκληματολογικά Πειστήρια.

1. ΕΙΣΑΓΩΓΗ

Στο Διαδίκτυο σήμερα εκτυλίσσεται ένα μεγάλο μέρος των καθημερινών ανθρώπινων δραστηριοτήτων, όπως της εργασίας, της διασκέδασης, της εκπαίδευσης και της κοινωνικής δικτύωσης. Εκτός από την αύξηση στο πλήθος των υπηρεσιών που προσφέρονται, αλλά και τη διεύρυνση της ποικιλίας τους, το πλήθος των χρηστών που συνδέονται αυξάνεται με ρυθμούς πολύ μεγαλύτερους από αυτούς του παγκόσμιου πληθυσμού, με τους αριθμούς να μιλάνε για 3,5 δισεκατομμύρια χρήστες του Διαδικτύου.

Παράλληλα, αύξηση της ποικιλομορφίας παρουσιάζεται και στους τρόπους με τους οποίους οι χρήστες μπορούν να επικοινωνήσουν μεταξύ τους. Έτσι, για παράδειγμα, υπάρχει η δυνατότητα επικοινωνίας με άμεσα ή έμμεσα γραπτά μηνύματα, με μετάδοση φωνής, με βίντεο, με διαμοιραζόμενα μέσα, ή και με άλλους

τρόπους. Παρότι όμως τα μέσα επικοινωνίας γίνονται περισσότερα, αλλά και πιο εξελιγμένα, και παρότι οι ταχύτητες σύνδεσης γίνονται υψηλότερες, με τεχνολογίες που προσφέρουν διάφορους τρόπους πρόσβασης, ενσύρματης ή ασύρματης, μελέτες και στατιστικές (Internet live stats, n.d.) δείχνουν ότι το γραπτό κείμενο παραμένει το κύριο μέσο επικοινωνίας μεταξύ των χρηστών του Διαδικτύου. Η κύρια συσκευή εισόδου, με την οποία παράγεται κείμενο, είναι το πληκτρολόγιο QWERTY. Αν και προτάθηκαν διάφορες διατάξεις πληκτρολογίου, μερικές από τις οποίες συναντώνται και σήμερα σε διάφορες ηλεκτρονικές συσκευές, το πληκτρολόγιο QWERTY κατέχει τη μερίδα του λέοντος στις συσκευές σύνταξης κειμένου. Μάλιστα, εκτός από την παραδοσιακή μορφή του, ως αναπόσπαστο κομμάτι των desktops και των laptops, σήμερα έχει και εικονική μορφή στις πιο «σύγχρονες» και περισσότερο φορητές συσκευές, τα tablets και τα smartphones.

Εκτός όμως από την πληθώρα των υπηρεσιών του Διαδικτύου, παρουσιάζεται και μία ιδιομορφία στον τρόπο επικοινωνίας μεταξύ των χρηστών. Αυτή είναι η δυνατότητα κάποιου να παραμείνει ανώνυμος, είτε αποκρύπτοντας τα δηλωτικά χαρακτηριστικά του, είτε υιοθετώντας κάποια αναληθή. Η ανωνυμία αυτή είναι ως ένα σημείο επιθυμητή αφού παρέχει κάποια αίσθηση ελευθερίας στους χρήστες και κάποια διασφάλιση ότι δεν θα διαρρεύσουν προσωπικά ή ευαίσθητα δεδομένα τους. Πολλές φορές όμως αποτελεί πρόβλημα ή πηγή προβλημάτων.

Για παράδειγμα, όταν ο χρήστης αποκρύπτει την ταυτότητά του, δεν επωφελείται από υπηρεσίες του Διαδικτύου που εστιάζουν στα χαρακτηριστικά του και που μπορεί να του πρότειναν να επισκεφτεί συγκεκριμένους Δικτυακούς Τόπους του ενδιαφέροντός του, ή να συμμετέχει σε συζητήσεις των προτιμήσεών του, ή να ενταχθεί σε ομάδες με μέλη που έχουν κοινά χαρακτηριστικά με αυτόν.

Σημαντικότερο όμως είναι ότι η ανωνυμία τροποποιεί τη συμπεριφορά ενός χρήστη (Suler, 2004), αφού αίρονται κάποιες αναστολές του και πιθανόν, πέρα από την επιθυμητή απελευθέρωση της ευπρεπούς έκφρασης, να οδηγείται και σε παράνομες ενέργειες, τις οποίες δεν θα διέπραττε εάν η ταυτότητά του ήταν γνωστή. Επιπλέον, η ανωνυμία αποτελεί το μεγαλύτερο πλεονέκτημα των παράνομων χρηστών, οι οποίοι δίνοντας ψεύτικες πληροφορίες για την ταυτότητά τους, προσπαθούν να αποκτήσουν την εμπιστοσύνη ανυποψίαστων χρηστών με σκοπό την εκμετάλλευση ή βλάβη τους. Οικονομικές απάτες, αποπλάνηση ανηλίκων, σχολικός εκφοβισμός, συκοφαντία, διανομή παιδικής πορνογραφίας, απειλές, διανομή ιών υπολογιστών, κ.α., είναι ορισμένες από τις κακόβουλες ενέργειες που βασίζονται στην ανωνυμία.

Άρα, η αποκάλυψη κάποιων εγγενών ή επίκτητων χαρακτηριστικών ενός χρήστη, όπως το φύλο, η ηλικία, το επικρατές χέρι, το μορφωτικό επίπεδο, η μητρική γλώσσα, κ.α., τα οποία ηθελημένα ή από αμέλεια δεν δήλωσε κατά τη συνομιλία του με άλλον χρήστη ή κατά τη χρήση κάποιων Διαδικτυακών υπηρεσιών, θα ήταν ωφέλιμη για την καλύτερη εκμετάλλευση των δυνατοτήτων του Διαδικτύου, για την ενημέρωση ανυποψίαστων χρηστών για ενδεχόμενους κινδύνους, αλλά και για την παροχή χρήσιμων πληροφοριών, εγκληματολογικού ενδιαφέροντος, στις περιπτώσεις όπου κάποιο ηλεκτρονικό αδίκημα έχει διαπραχθεί.

Στην παρούσα εργασία επιχειρείται η ταξινόμηση χρηστών με σκοπό τον εντοπισμό κάποιων χαρακτηριστικών τους. Αυτό επιτυγχάνεται με τη βοήθεια της δυναμικής της πληκτρολόγησης (keystroke dynamics), δηλαδή, του τρόπου με τον οποίο ένας χρήστης πληκτρολογεί. Ως παράμετροι χρησιμοποιούνται διάρκειες πατήματος πλήκτρου (keystroke durations) και λανθάνοντες χρόνοι διγράμματος DDDL (down-down digram latency), ενώ το χαρακτηριστικό που αναζητείται είναι η ημερήσια χρήση της ηλεκτρονικής τους συσκευής. Η γνώση του πλήθους των ωρών ανά ημέρα που ξοδεύει ένας χρήστης σε υπολογιστή δεν μπορεί να αποτελέσει ασφαλές συμπέρασμα για τα χαρακτηριστικά του, αλλά μπορεί να οδηγήσει σε κάποιες υποθέσεις. Για παράδειγμα, να καταδείξει την ηλικία του, αφού σύμφωνα με στατιστικές μελέτες (Daily computer usage in Great Britain by age 2006-2015, n.d.) αυτά τα δύο μεγέθη συνδέονται αντιστρόφως ανάλογα, το φύλο του, καθώς οι άντρες καταναλώνουν περισσότερο χρόνο στον υπολογιστή από ότι οι γυναίκες (Winn & Heeter, 2009), και το ετήσιο εισόδημά του, εφόσον φαίνεται ότι τα υψηλότερα εισοδήματα αντιστοιχίζονται σε λιγότερες ώρες μπροστά σε υπολογιστή (Hofferth et al., 2013). Ακόμα, η γνώση αυτή μπορεί να βοηθήσει στον αποκλεισμό υπόπτων στην περίπτωση εγκληματολογικής έρευνας, αφού μάλλον θα ήταν απίθανο ένα άτομο με ελάχιστες ώρες ενασχόλησης να διαπράξει μία κυβερνοεπίθεση, ή κάτι παραπλήσιο.

2. ΑΝΑΣΚΟΠΗΣΗ ΒΙΒΛΙΟΓΡΑΦΙΑΣ

Για την επίλυση του προβλήματος της άρσης της πλήρους ανωνυμίας στο Διαδίκτυο έχουν προταθεί διάφορες μέθοδοι. Μία από αυτές είναι η εκμετάλλευση πληροφοριών από φωτογραφίες προσώπου που μπορεί ένας χρήστης να έχει στο προφίλ του. Έτσι οι Alrashed και Berbar (2013) σχεδίασαν σύστημα βασιζόμενο σε ταξινομητή μηχανής υποστήριξης διανυσμάτων (support vector machine, SVM) και εξετάζοντας 1982 φωτογραφίες πέτυχαν ένα ποσοστό ορθής πρόβλεψης φύλου κοντά στο 99,5%. Για την ακρίβεια των αποτελεσμάτων τους χρησιμοποίησαν τα στατιστικά μεγέθη ευαισθησία (sensitivity), το οποίο υπολογίζει το ποσοστό των αληθώς θετικών (true positive rate) αποτελεσμάτων, και ειδικότητα (specificity), το οποίο υπολογίζει το ποσοστό των αληθώς αρνητικών (true negative rate) αποτελεσμάτων. Επίσης, οι Damayanti και Rachmad (2016) σχεδίασαν παρόμοιο σύστημα, με την ακρίβειά του να κυμαίνεται από 74% έως 92%. Στο πεδίο της αναζήτησης της ηλικίας ενός χρήστη από φωτογραφίες προσώπου, οι Hewahi κ.ά. (2010) όρισαν 4 ηλικιακές ομάδες, επέλεξαν 68 σημεία πάνω σε μία εικόνα προσώπου και πέτυχαν ποσοστό ορθής πρόβλεψης 78,4%. Ο Choobeh (2012) με τον ίδιο αριθμό σημείων, επέλεξε 130 παραμέτρους, τις οποίες εφάρμοσε σε καθένα από τα νευρωνικά δίκτυα που ενεπλάκησαν στη διαδικασία πετυχαίνοντας ένα μέσο απόλυτο σφάλμα (mean absolute error) της ακριβούς ηλικίας του χρήστη μεταξύ 4,85 και 5,85 έτη. Ενώ, οι Yi κ.ά. (2014), χρησιμοποιώντας ένα συνελικτικό νευρωνικό δίκτυο (convolutional neural network, CNN), μείωσαν περαιτέρω το μέσο απόλυτο σφάλμα στα 3,63 έτη, βελτιώνοντας ταυτόχρονα και την ταχύτητα του συστήματος.

Μία άλλη μέθοδος που αφορά την εξεύρεση κάποιων χαρακτηριστικών των χρηστών, είναι εξέταση των κειμένων που παράγουν. Όπως η προσπάθεια των

Mukherjee και Liu (2010) για να κατατάξουν τους συγγραφείς των blogs σύμφωνα με το φύλο τους. Το σύστημά τους υλοποιήθηκε και με ταξινομητή Naïve Bayes, αλλά και με SVM, επιτυγχάνοντας ποσοστό ορθής πρόβλεψης της τάξης του 88%. Για την επιλογή των κατάλληλων παραμέτρων (features selection) που θα τους οδηγούσαν στα επιθυμητά αποτελέσματα, οι ερευνητές χρησιμοποίησαν το στατιστικό χ^2 (chi square statistic) που μετρά την έλλειψη ανεξαρτησίας μεταξύ μιας παραμέτρου και μιας κλάσης. Επίσης, οι Cheng κ.ά. (2011) προσπάθησαν να λύσουν το ίδιο πρόβλημα με δεδομένα προερχόμενα από μία συλλογή κειμένων από ομάδες συζητήσεων του Reuters και μια συλλογή από e-mails των εργαζομένων της εταιρίας Enron. Μελέτησαν 545 παραμέτρους, εκπαίδευσαν τρεις ταξινομητές και τα αποτελέσματα έδειξαν τον SVM ως πιο αποδοτικό, το σύστημα να βελτιώνεται όσο αύξανε το σύνολο εκπαίδευσης (training set) και όσο ελέγχονταν μεγαλύτερα κείμενα, με ένα μέγιστο ποσοστό επιτυχίας της τάξης του 85%. Μάλιστα, σε μια διαδικασία μείωσης του πλήθους των παραμέτρων, έτσι ώστε να μπορέσουν να κάνουν το σύστημά τους ταχύτερο, οι ερευνητές αξιολόγησαν τη σημαντικότητα των παραμέτρων χρησιμοποιώντας τη στατιστική δοκιμή t-test. Θέτοντας το επίπεδο σημαντικότητας στο 5%, κατέληξαν σε 157 παραμέτρους, με τη χρήση των οποίων το σύστημα γινόταν 3 φορές ταχύτερο, με απώλεια 3% στην ακρίβεια.

Η αναζήτηση του φύλου και της ηλικίας του δημιουργού ενός blog ήταν το αντικείμενο μελέτης της εργασίας των Schler κ.ά. (2006). Οι ερευνητές όρισαν τρεις κλάσεις, την 10's που αντιστοιχούσε στις ηλικίες 13-17, την 20's που αντιστοιχούσε στις 23-27, και στην 30's που αντιστοιχούσε στις 33-42. Για την ταξινόμηση επιστρατεύτηκε ο αλγόριθμος Multi-Class Real Winnow και τα τελικά αποτελέσματα έδειξαν ότι η ηλικιακή ομάδα μπορούσε να προβλεφθεί ορθώς με ένα ποσοστό της τάξης του 73%, ενώ το φύλο κοντά στο 80%.

Τέλος, τα χαρακτηριστικά των χρηστών αναζητήθηκαν με παραμέτρους που προέκυψαν από τη συμπεριφορά τους σε κοινωνικά δίκτυα. Για παράδειγμα, στην εργασία των Rao κ.ά. (2010) επιχειρείτε η αναγνώριση χαρακτηριστικών των χρηστών του Twitter. Μεταξύ των πεδίων στα οποία αναζητήθηκαν παράμετροι ήταν η δομή του κοινωνικού τους δικτύου και η κοινωνική συμπεριφορά τους. Οι χρήστες χωρίστηκαν σε δύο ηλικιακές ομάδες και οι ερευνητές δοκίμασαν το σύστημά τους για τις παραμέτρους καθενός από τα πεδία που εντόπισαν, πετυχαίνοντας ποσοστό ορθής πρόβλεψης κοντά στο 74%.

Ωστόσο, όλες οι παραπάνω μέθοδοι αντιμετωπίζουν κάποιους περιορισμούς στη γενίκευσή τους. Ο λόγος είναι ότι απαιτούνται συγκεκριμένα δεδομένα (π.χ. φωτογραφίες προσώπου) από πλευράς χρήστη, ή υπάρχει εξάρτηση από συγκεκριμένη ομιλούμενη γλώσσα, αφού οι παράμετροι προέρχονται από συγκεκριμένες φράσεις, λέξεις και N-γράμματα, ή θεωρείται δεδομένη μια φυσιολογική συμπεριφορά του χρήστη, ο οποίος πρέπει να διατηρεί λογαριασμό σε κάποιο κοινωνικό δίκτυο. Στην παρούσα εργασία προτείνεται ο εντοπισμός κάποιων εγγενών ή επίκτητων χαρακτηριστικών χρήστη δια μέσω της δυναμικής της πληκτρολόγησης, που ορίζεται ως η λεπτομερής χρονική καταγραφή των ενεργειών του στο πληκτρολόγιο, δηλαδή το πότε πίεσε και το πότε απελευθέρωσε κάθε πλήκτρο. Η μέθοδος αυτή, εκτός του ότι δεν απαιτεί παρά τα πιο απλά δεδομένα που

παράγει ο χρήστης και εκτός του ότι είναι ανεξάρτητη από την ομιλούμενη γλώσσα που χρησιμοποιεί, δεν είναι παρεμβατική, με την έννοια ότι δεν είναι αναγκαία η ανάγνωση των κειμένων του, η παρακολούθηση της συμπεριφοράς του στο Διαδίκτυο και η αναζήτηση των συνδέσεων του κοινωνικού του δικτύου.

Η δυναμική της πληκτρολόγησης χρησιμοποιήθηκε ως επί το πλείστον στην αυθεντικοποίηση χρηστών, με σκοπό την αντικατάσταση του παραδοσιακού σχήματος με τη χρήση του κωδικού πρόσβασης. Όπως για παράδειγμα στην εργασία των Patil και Renke (2016), οι οποίοι κατέγραψαν χρήστες μιας εφαρμογής, αποθηκεύοντας τα δεδομένα σε βάση δεδομένων, και τα οποία στη συνέχεια τα χρησιμοποιούσαν για να επαληθεύσουν εάν ο χρήστης που πληκτρολογεί είναι αυτός που ισχυρίζεται ότι είναι. Επίσης, οι Wankhede και Verma (2014) πέτυχαν ποσοστό εσφαλμένης απόρριψης (false rejection rate, FRR) 4,8 % και ποσοστό εσφαλμένης αποδοχής (false acceptance rate, FAR) 3,1% στο σύστημα αυθεντικοποίησής τους, που υλοποιήθηκε με πολυστρωματικό perceptron (multi-layer perceptron, MLP). Μάλιστα, με σκοπό την απομάκρυνση δεδομένων που αύξαναν το βαθμό ασυνέπειας των χρηστών στην πληκτρολόγηση, χρησιμοποίησαν το στατιστικό μέγεθος Z-score, το οποίο υπολογίζεται με τη βοήθεια της μέσης τιμής και της τυπικής απόκλισης των τιμών κάθε παραμέτρου.

Από τη δυναμική της πληκτρολόγησης μπορούν να εξαχθούν πολλές παράμετροι, κάθε μία από τις οποίες περικλείει μικρή ποσότητα πληροφορίας, ο συνδυασμός τους όμως είναι ικανός να δώσει ικανοποιητικά αποτελέσματα, όπως τουλάχιστον φάνηκε από σχετικές μελέτες,. Έτσι, χρησιμοποιείται η διάρκεια πατήματος πλήκτρου (keystroke duration) που ορίζεται ως ο χρόνος που πέρασε από τη στιγμή που ένα πλήκτρο πατήθηκε, μέχρι τη στιγμή που ελευθερώθηκε. Μια άλλη παράμετρος είναι ο λανθάνων χρόνος διγράμματος (digram latency) που ορίζεται ως ο χρόνος που χρειάστηκε ένας χρήστης για να χρησιμοποιήσει δύο συνεχόμενα πλήκτρα. Το μέγεθος αυτό μπορεί να εκφραστεί με τέσσερις διαφορετικούς τρόπους (Hosseinzadeh & Krishnan, 2008), που προκύπτουν από τους συνδυασμούς πίεσης και απελευθέρωσης των δύο πλήκτρων, και συγκεκριμένα είναι τα down-down digram latency (DDDL), up-up digram latency (UUDL), down-up digram latency (DUDL) και up-down digram latency (UDDL). Με παρόμοιους τρόπους ορίζονται και ο λανθάνων χρόνος τριγράμματος (trigram latency), ο λανθάνων χρόνος τετραγράμματος (tetragram latency) και γενικά ο λανθάνων χρόνος N-γράμματος (N-gram latency) (Zhao, 2006).

Εκτός όμως από τις παραμέτρους που σχετίζονται με το χρόνο, στις μελέτες που διεξήχθησαν και αφορούσαν τη δυναμική της πληκτρολόγησης, εντάσσονται και άλλες παράμετροι που δεν σχετίζονται με τη λεπτομερή καταγραφή των χρόνων των συμβάντων. Ως τέτοιες λογίζονται η ταχύτητα πληκτρολόγησης (λέξεις ανά λεπτό), η συχνότητα λαθών κατά την πληκτρολόγηση, ο τρόπος διόρθωσης λαθών, το ποσοστό χρήσης πλήκτρων που συναντώνται παραπάνω από μία φορές στο πληκτρολόγιο (όπως το “Shift” (Bartlow & Cukic, 2006), το “Ctrl”, το “Alt”, το “Enter”, κτλ) (Kumar et al, 2014), η ώρα της ημέρας που κάποιος χρήστης επιλέγει να πληκτρολογήσει, οι εφαρμογές στις οποίες πληκτρολογεί και γενικώς η συχνότητα χρήσης του πληκτρολογίου.

3. ΜΕΘΟΔΟΛΟΓΙΑ

Η μέθοδος που ακολουθήθηκε αποτελείται από τρία διακριτά στάδια. Στο πρώτο στάδιο έγινε η λήψη των απαραίτητων δεδομένων δυναμικής της πληκτρολόγησης από εθελοντές χρήστες. Στο δεύτερο στάδιο έγινε η επιλογή των παραμέτρων που χρησιμοποιήθηκαν για την κατηγοριοποίηση των χρηστών. Τέλος, στο τρίτο στάδιο έγινε η επιλογή που κατάλληλου ταξινομητή, έτσι ώστε να προκύψει το σύστημα με τη βέλτιστη απόδοση, σε ότι αφορά το ποσοστό ορθής πρόβλεψης, την ταχύτητα λειτουργίας και την σταθερότητα στην εξαγωγή αποτελεσμάτων.

3.1 Λήψη Δεδομένων

Τα δεδομένα που λαμβάνονται από τη δυναμική της πληκτρολόγησης μπορεί να προέρχονται από καθορισμένο κείμενο (fixed) ή από ελεύθερο κείμενο (free text). Ως ελεύθερο κείμενο νοείται ένα συγκεκριμένο κείμενο που έχει δοθεί σε έναν εθελοντή χρήστη να πληκτρολογήσει ενώ βρίσκεται σε κατάσταση καταγραφής της πληκτρολόγησής του. Συνήθως το καθορισμένο κείμενο πληκτρολογείται σε κάποιο κλειστό περιβάλλον. Ως ελεύθερο κείμενο νοείται το κείμενο που πληκτρολογεί ο εθελοντής χρήστης κατά βούληση ενώ βρίσκεται σε κατάσταση καταγραφής της πληκτρολόγησής του. Καταγραφή πληκτρολόγησης ελεύθερου κειμένου μπορεί να διεξάγεται σε κλειστό περιβάλλον ή κατά τη διάρκεια καθημερινής χρήσης του υπολογιστή, όπου ο χρήστης χρησιμοποιεί τις εφαρμογές που επιθυμεί και πληκτρολογεί τις χρονικές στιγμές που επιθυμεί.

Στη συγκεκριμένη έρευνα, για λόγους που έχουν να κάνουν με την όσο το δυνατόν πλησιέστερη αναπαράσταση των πραγματικών συνθηκών, αλλά και για την καταγραφή παραμέτρων που δεν βρίσκονταν στον αρχικό σχεδιασμό, επιλέχθηκε η λήψη δεδομένων από ελεύθερο κείμενο. Όμως, τα διαθέσιμα σύνολα δεδομένων (datasets) της δυναμικής της πληκτρολόγησης στο Διαδίκτυο είναι ελάχιστα, με αυτά που προέρχονται από ελεύθερο κείμενο να είναι η μειονότητα. Μάλιστα, δεν φαίνεται να υπάρχει ούτε ένα σύνολο δεδομένων ελεύθερου κειμένου που να προήλθε από την καταγραφή μεγάλου κειμένου, τουλάχιστον σύμφωνα με όσα είναι γνωστά,. Τα όσα είναι διαθέσιμα περιορίζονται στην αποτύπωση ορισμένων λέξεων ή φράσεων, με τον λόγο να είναι προφανής και να μην είναι άλλος από το ότι αυτά τα δεδομένα μπορούν να αποκαλύψουν κωδικούς πρόσβασης, αριθμούς πιστωτικών καρτών, προσωπικά μηνύματα, και άλλες ευαίσθητες πληροφορίες του καταγεγραμμένου εθελοντή.

Για το λόγο αυτό αποφασίστηκε να δημιουργηθεί ένα νέο σύνολο δεδομένων για τις ανάγκες αυτής της έρευνας. Για την εκπλήρωση αυτού του στόχου σχεδιάστηκε λογισμικό καταγραφής πληκτρολόγησης (keylogger), με το όνομα «IRecU», το οποίο εγκαταστάθηκε στις συσκευές των εθελοντών. Κατά την πρώτη είσοδο του χρήστη στο «IRecU» του ζητούταν να συμπληρώσει μία φόρμα με ορισμένα χαρακτηριστικά του, ανάμεσα στα οποία ήταν η κατά μέσο όρο ημερήσια χρήση της ηλεκτρονικής συσκευής του. Οι επιλογές που δόθηκαν στους χρήστες ήταν 5, οι «0-1 ώρες», «1-2 ώρες», «2-4 ώρες», «4-6 ώρες» και «6+ ώρες», με συνέπεια να δημιουργηθούν 5

κλάσεις. Μετά το πέρας της διαδικασίας καταγραφής, η οποία διήρκεσε από 20/02/2014 έως και 27/12/2014, συλλέχθηκαν 248 αρχεία από 75 εθελοντές χρήστες. Κάθε αρχείο περιέχει δεδομένα από 2.800 έως 4.500 πατήματα πλήκτρων, καταγεγραμμένα σε εγγραφές της μορφής:

78,#2014-06-20#,34680537,"dn"
78,#2014-06-20#,34680657,"up"
65,#2014-06-20#,34680687,"dn"
73,#2014-06-20#,34680787,"dn"
65,#2014-06-20#,34680797,"up"
73,#2014-06-20#,34680887,"up"

Τα πεδία σε κάθε εγγραφή χωρίζονται με κόμμα (,). Στο πρώτο πεδίο δηλώνεται το virtual key code του πλήκτρου στο οποίο έγινε η ενέργεια, σε δεκαδική μορφή. Στο δεύτερο πεδίο, ανάμεσα στα σύμβολα της δίεσης (#), δηλώνεται η ημερομηνία που έγινε η ενέργεια πληκτρολόγησης. Στο τρίτο πεδίο δηλώνονται τα ms που πέρασαν από την αρχή της συγκεκριμένης ημέρας, τη στιγμή που έγινε η ενέργεια. Τέλος, στο τέταρτο πεδίο δηλώνεται το είδος της ενέργειας, με "dn" να αντιστοιχεί στην πίεση πλήκτρου και με "up" στην απελευθέρωση πλήκτρου.

Το διαθέσιμο πλήθος αρχείων ανά κλάση και το ποσοστό τους επί του συνόλου, παρουσιάζεται στον Πίνακα 1.

Πίνακας 1. *Πλήθος και ποσοστό αρχείων ανά κλάση*

	Πλήθος	**Ποσοστό**
Αρχεία 0-1 ωρών	23	9,3%
Αρχεία 1-2 ωρών	46	18,5%
Αρχεία 2-4 ωρών	54	21,8%
Αρχεία 4-6 ωρών	38	15,3%
Αρχεία 6 και άνω ωρών	87	35,1%
Συνολικά αρχεία	248	100,0%

Αν και το σύνολο δεδομένων δεν είναι ισορροπημένο, η αντιπροσώπευση κάθε κλάσης είναι επαρκής, με συνέπεια να θεωρούνται αξιόπιστα τα αποτελέσματα.

3.2 Εξαγωγή Παραμέτρων

Όπως προαναφέρθηκε, η δυναμική της πληκτρολόγησης συνοδεύεται από μεγάλο πλήθος παραμέτρων, με τις περισσότερες έρευνες να χρησιμοποιούν τις διάρκειες πατήματος πλήκτρου και τους λανθάνοντες χρόνους διγράμματος. Κάποιοι (Douhou & Magnus, 2009) ισχυρίζονται ότι η χρήση των keystroke durations φέρνει καλύτερα αποτελέσματα, ενώ κάποιοι άλλοι (Hassan et al., 2013) ότι οι λανθάνοντες χρόνοι διγράμματος είναι προτιμότεροι. Για το λόγο αυτό, όπως ήδη ειπώθηκε, σε αυτή την

εργασία χρησιμοποιούνται και τα δύο είδη παραμέτρων. Μάλιστα, για την αποφυγή αρνητικών τιμών, χρησιμοποιούνται οι λανθάνοντες χρόνοι διγράμματος DDDL.

Όμως, θεωρώντας ένα πληκτρολόγιο των 100 πλήκτρων, τότε από τη χρήση του μπορούν να εξαχθούν 100 keystroke durations και περίπου 400.000 digram latencies, εκ των οποίων οι περίπου 100.000 είναι DDDL. Αυτός είναι ένα πολύ μεγάλος αριθμός παραμέτρων που θα οδηγήσει σε συστήματα τα οποία θα απαιτούν μεγάλο χρόνο εκπαίδευσης, αλλά και εξαγωγής αποτελεσμάτων, με συνέπεια να καθίστανται ανεπαρκή στις περιπτώσεις όπου είναι αναγκαία η άμεση λήψη απόφασης. Για το λόγο αυτό ήταν επιτακτικό να χρησιμοποιηθεί μόνο ένα υποσύνολο από τις διαθέσιμες παραμέτρους. Έτσι, ύστερα από έλεγχο των αρχείων καταγραφής, εντοπίστηκαν τα πλήκτρα και τα διγράμματα τα οποία χρησιμοποιήθηκαν σχεδόν από όλους τους χρήστες και με σχετικά μεγάλη συχνότητα. Η διαδικασία αυτή κατέληξε σε 60 διάρκειες πατήματος πλήκτρου και 140 λανθάνοντες χρόνους διγράμματος.

Για την εξαγωγή των παραμέτρων σχεδιάστηκε νέο λογισμικό, με το όνομα «ISqueezeU», το οποίο δεχόταν ως είσοδο αρχεία κειμένου μορφής όπως αυτά που παραγόταν από το «IRecU» και εξήγαγε τη μέση τιμή των παραμέτρων που είχαν προεπιλεγεί, εφόσον το πλήθος εμφανίσεων του αντίστοιχου πλήκτρου ή διγράμματος υπερέβαινε την τιμή ενός κατωφλίου που επίσης είχε προκαθοριστεί. Το κατώφλι ορίστηκε στις 10 εμφανίσεις για τα πλήκτρα και στις 5 εμφανίσεις για τα διγράμματα, ενώ ο λόγος ύπαρξής του ήταν η απομάκρυνση των τιμών που δεν θα ήταν αντιπροσωπευτικές της συμπεριφοράς ενός χρήστη επί του πληκτρολογίου.

Ένα παράδειγμα της εξόδου του «ISqueezeU» παρουσιάζεται στον Πίνακα 2.

Πίνακας 2. *Παράδειγμα εξόδου του «ISqueezeU»*

Παράμ.	**Αρχείο Καταγραφής**							
	003	**113**	**127**	**155**	**198**	**209**	**218**	**227**
32	96,2	64,1	126,0	92,6	136,6	126,5	77,5	83,3
65	82,2	91,0	120,1	85,6	152,3	162,5	50,2	72,6
77	87,6	66,5	121,4	71,1	166,3	139,6	62,3	74,1
87	?	80,8	94,4	77,8	134,9	?	58,4	62,5
32-68	562,6	448,4	881,1	544,0	?	1169,8	404,0	667,8
65-73	277,1	115,5	247,2	227,2	158,0	299,3	113,6	156,0
71-82	249,9	296,3	?	303,0	148,4	211,8	284,6	180,6
77-69	356,1	145,9	222,6	236,6	190,2	356,2	198,8	145,6

Το λατινικό ερωτηματικό (?) δηλώνει ότι η συγκεκριμένη παράμετρος, στο συγκεκριμένο αρχείο καταγραφής, δεν είχε επαρκή αριθμό εμφανίσεων.

3.3 Ταξινομητής

Για την επίτευξη της καλύτερης απόδοσης του συστήματος δοκιμάστηκε ένα πλήθος ταξινομητών, συμπεριλαμβανομένων αυτών που βασίζονται σε απόσταση

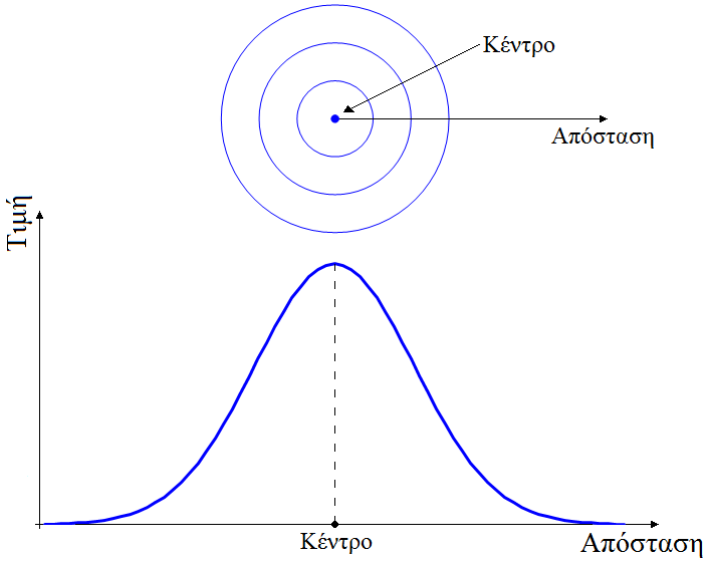
(Ευκλείδεια, Manhattan, κτλ), σε δέντρα απόφασης (decision trees), στο θεώρημα του Bayes, σε μηχανές υποστήριξης διανυσμάτων και σε νευρωνικά δίκτυα.

Η κατάληξη αυτής της διαδικασίας ήταν ένα νευρωνικό δίκτυο με συνάρτηση ακτινωτής βάσης (radial basis function network, RBFN), το οποίο επιλέχτηκε γιατί παρουσίαζε υψηλό ποσοστό πρόβλεψης και μικρό χρόνο εκπαίδευσης (training time).

Τα νευρωνικά δίκτυα με συνάρτηση ακτινωτής βάσης διατυπώθηκαν για πρώτη φορά από τους Broomhead και Lowe (1988) και μία από τις σημαντικές διαφορές που παρουσιάζουν, σε σχέση με έναν MLP, είναι ότι ως συνάρτηση μεταφοράς χρησιμοποιούν συνάρτηση ακτινωτής βάσης (radial basis function, RBF), από όπου πήραν και το όνομά τους. Η τιμή που επιστρέφει μία RBF εξαρτάται μόνο από την απόσταση της μεταβλητής από ένα σημείο, το οποίο ονομάζεται κέντρο. Όταν αυτή η απόσταση είναι μηδενική, τότε η συνάρτηση παίρνει τη μέγιστη τιμής της, ενώ όταν η απόσταση τείνει στο άπειρο, τότε η τιμή της τείνει στο μηδέν.

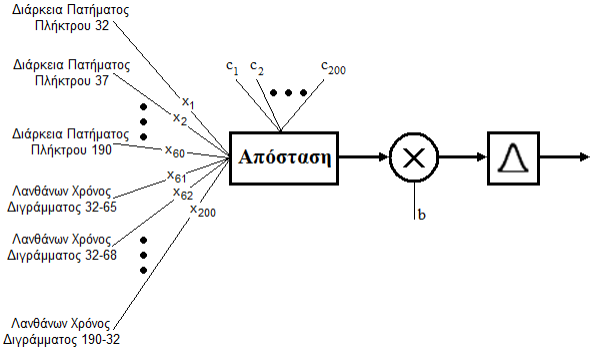
Η μορφή μιας RBF είναι παραπλήσια με αυτή του Σχήματος 1.

Σχήμα 1. *Συνάρτηση ακτινωτής βάσης*



Επίσης, η λειτουργία ενός νευρώνα RBF, η οποία αναπαριστάται στο Σχήμα 2, είναι διαφορετική από αυτή ενός perceptron.

Σχήμα 2. *Λειτουργία νευρώνα δικτύου με συνάρτηση ακτινωτής βάσης*



Οι είσοδοι του νευρωνικού δικτύου x_1, x_2, …, x_{200}, των 60 διαρκειών πατήματος πλήκτρου και των 140 λανθανόντων χρόνων διγράμματος, σχηματίζουν ένα διάνυσμα *x* 200 διαστάσεων. Το διάνυσμα αυτό εφαρμόζεται στην είσοδο του νευρώνα και υπολογίζεται η απόστασή του από ένα άλλο διάνυσμα *c*, ίδιων διαστάσεων, που είναι το διάνυσμα-κέντρο του νευρώνα. Η απόσταση μεταξύ των δύο διανυσμάτων, που συμβολίζεται ||*x*-*c*||, τυπικά λαμβάνεται από τον υπολογισμό της Ευκλείδειας απόστασης, αν και η χρησιμοποίηση της απόστασης Mahalanobis φαίνεται να αποδίδει καλύτερα.

Η υπολογισμένη απόσταση πολλαπλασιάζεται με έναν συντελεστή *b* και στο γινόμενο εφαρμόζεται η συνάρτηση ακτινωτής βάσης. Το αποτέλεσμα της συνάρτησης αποτελεί και την έξοδο του νευρώνα που γράφεται ως:

$$y_i(x) = r(\|x - c\|) \tag{1}$$

Η συνάρτηση *r* είναι ακτινωτής βάσης και επιλέγεται να δίνεται από τον τύπο:

$$r(\|x - c\|) = e^{-b\cdot\|x-c\|^2} \tag{2}$$

Ο δείκτης *i* στην εξίσωση (1) δηλώνει ότι πρόκειται για την έξοδο του *i* νευρώνα του δικτύου, αφού για την επίλυση ενός προβλήματος όπως η κατηγοριοποίηση χρηστών ανά ημερήσια χρήση υπολογιστή, με παραμέτρους τα keystroke durations και τα digram latencies, απαιτείται η χρησιμοποίηση πολλών νευρώνων που σχηματίζουν ένα νευρωνικό δίκτυο με συνάρτηση ακτινωτής βάσης.

Ο συνδυασμός των εξόδων των νευρώνων για τον υπολογισμό της τελικής εξόδου του δικτύου γίνεται με γραμμικό τρόπο, με τη βοήθεια συντελεστών *α*:

$$y(x) = \sum_{i=1}^{N} \alpha_i \cdot r(\|x - c_i\|) \tag{3}$$

Όπου *N* είναι το πλήθος των νευρώνων του δικτύου στο κρυμμένο στρώμα (hidden layer) και όπου c_i είναι το διάνυσμα-κέντρο του *i* νευρώνα.

Από τις εξισώσεις (2) και (3) προκύπτει η έξοδος του νευρωνικού δικτύου:

$$y(x) = \sum_{i=1}^{N} \alpha_i \cdot e^{-b_i \cdot \|x - c_i\|^2} \quad (4)$$

Οι συντελεστές α_i και b_i, καθώς και τα διανύσματα c_i, λαμβάνουν τέτοιες τιμές ώστε η απόδοση του ταξινομητή να βελτιστοποιηθεί.

Συγκεκριμένα, κατά τη φάση της εκπαίδευσης του συστήματος, σχηματίζονται συστάδες (clusters) των δεδομένων εισόδου. Υπενθυμίζεται ότι οι τιμές των παραμέτρων της δυναμικής της πληκτρολόγησης του κάθε δείγματος, που εφαρμόζονται ως είσοδος, ορίζουν ένα διάνυσμα 200 διαστάσεων. Σε κάθε συστάδα δεδομένων εισόδου αντιστοιχίζεται ένα διάνυσμα-κέντρο, ίδιου αριθμού διαστάσεων. Ο διαχωρισμός των δειγμάτων σε συστάδες και κατά συνέπεια ο υπολογισμός των κέντρων τους, μπορεί να γίνει με οποιονδήποτε αλγόριθμο συσταδοποίησης (clustering algorithm), όπως για παράδειγμα με αυτόν των *k* μέσων (*k*-means clustering algorithm), που χρησιμοποιήθηκε για πρώτη φορά από τον MacQueen (1967). Το πλήθος των συστάδων αποτελεί μία παράμετρο σχεδιασμού του ταξινομητή, που καθορίζει το πλήθος των νευρώνων στο κρυμμένο στρώμα του δικτύου. Τα υπολογισμένα κέντρα των συστάδων γίνονται τα κέντρα των νευρώνων.

Ο σχηματισμός των συστάδων εξαρτάται και από μία ακόμα παράμετρο σχεδιασμού, την ελάχιστη τυπική απόκλιση (minimum standard deviation) που επιτρέπεται να έχουν τα σύνολα δεδομένων που τις απαρτίζουν. Όταν η ελάχιστη τυπική απόκλιση είναι αρκετά μικρή, τότε είναι πιθανό ο ταξινομητής να δημιουργήσει συστάδες που αποτελούνται από μία μόνο είσοδο.

4. ΠΕΙΡΑΜΑΤΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ

Αρχικώς, ελέγχθηκε η συμπεριφορά του συστήματος για διάφορες τιμές στις παραμέτρους του ταξινομητή. Συγκεκριμένα, μετρήθηκε το ποσοστό ορθής πρόβλεψης του συστήματος για τιμές στο πλήθος των συστάδων από 10 έως και 140, και για τιμές στην ελάχιστη επιτρεπτή τυπική απόκλιση από 1,0 έως και 1,9. Για να εξασφαλιστεί ότι τα αποτελέσματα των πειραμάτων είναι ανεξάρτητα από τα δεδομένα, ακολουθήθηκε η τεχνική της διασταυρωμένης επικύρωσης *k* φορών (*k*-fold cross-validation). Πρόκειται για μία τεχνική της στατιστικής επιστήμης, η οποία εδραιώθηκε στα τέλη της δεκαετίας του 1960, όταν αρκετές δημοσιεύσεις αναφερόντουσαν σε αυτή, όπως για παράδειγμα η εργασία των Lachenbruch και Mickey (1968). Τα αποτελέσματα των πειραμάτων εμφανίζονται στον Πίνακα 3.

Τα συμπεράσματα που εξάγονται από τον Πίνακα 3 είναι, κατά πρώτον, ότι το σύστημα παρουσιάζει ένα ποσοστό επιτυχίας 68%-75% για ένα μεγάλο εύρος των τιμών των δύο παραμέτρων σχεδίασης του ταξινομητή που μελετήθηκαν. Κατά δεύτερον, ότι το σύστημα έχει την ίδια συμπεριφορά, για την ίδια τιμή ελάχιστης τυπικής απόκλισης, όταν το πλήθος των συστάδων είναι πάνω από 80. Επίσης, κάτι που δεν παρουσιάζεται σε αυτόν τον Πίνακα, είναι ότι όταν το πλήθος των συστάδων μικραίνει, αυξάνεται ο χρόνος εκπαίδευσης του ταξινομητή. Ο λόγος είναι ότι όταν ο αριθμός συστάδων είναι μεγάλος, τότε αρκετές από αυτές αποτελούνται από ένα μόνο δείγμα, με αποτέλεσμα να μην δημιουργούνται πολλοί συνδυασμοί για δοκιμή.

Αντίθετα, όταν ο αριθμός των συστάδων είναι μικρός, υπάρχουν πολλοί διαφορετικοί συνδυασμοί που μπορούν να υλοποιήσουν τη συσταδοποίηση, και άρα ο χρόνος εκπαίδευσης είναι μεγάλος. Σε κάθε περίπτωση όμως, η λειτουργία του νευρωνικού δικτύου με συνάρτηση ακτινωτής βάσης είναι ταχύτερη αυτής του MLP.

Πίνακας 3. *Ποσοστό ορθής πρόβλεψης (%) για ζεύγη τιμών αριθμού συστάδων και ελάχιστης επιτρεπτής τυπικής απόκλισης, του ταξινομητή ημερήσιας χρήσης υπολογιστή*

		Πλήθος Συστάδων							
		10	**20**	**40**	**60**	**80**	**100**	**120**	**140**
Ελάχιστη Τυπική Απόκλιση	**1,0**	71,8	69,4	68,5	71,4	70,2	70,2	70,2	70,2
	1,1	70,6	68,2	70,9	74,2	71,8	71,8	71,8	71,8
	1,2	70,2	68,6	71,8	75,0	72,2	72,2	72,2	72,2
	1,3	69,0	70,2	73,0	74,6	72,2	72,2	72,2	72,2
	1,4	68,9	68,5	72,6	73,8	71,4	71,4	71,4	71,4
	1,5	68,6	70,1	73,0	73,4	70,2	70,2	70,2	70,2
	1,6	68,5	68,9	73,0	72,6	69,4	69,4	69,4	69,4
	1,7	66,5	68,9	71,8	72,2	67,7	67,7	67,7	67,7
	1,8	66,9	69,4	72,2	70,6	66,5	66,5	66,5	66,5
	1,9	66,1	70,2	72,6	69,0	64,1	64,1	64,1	64,1

Μια πιο εστιασμένη εικόνα των αποτελεσμάτων παρουσιάζεται στον Πίνακα 4, με τιμές παραμέτρων του ταξινομητή 100 συστάδες και 1,3 ελάχιστη επιτρεπτή τυπική απόκλιση.

Πίνακας 4. *Αποτελέσματα πρόβλεψης ημερήσιας χρήσης υπολογιστή*

	Προβλέφθηκαν ως					
Ημερήσια Χρήση Υπολογιστή	**0-1**	**1-2**	**2-4**	**4-6**	**6+**	**Ποσοστό Επιτυχίας ανά Κλάση**
0-1	16	2	0	1	4	69,6%
1-2	2	30	9	0	5	65,2%
2-4	3	1	39	3	8	72,2%
4-6	2	0	5	31	0	81,6%
6+	3	2	16	3	63	72,4%
Ποσοστό Επιτυχίας ανά Πρόβλεψη	61,5%	85,7%	56,5%	81,6%	78,8%	

Όπως φαίνεται, από το σύνολο των 248 αρχείων, τα 179 προβλέφθηκαν ορθώς, κάτι που σημαίνει ένα συνολικό ποσοστό επιτυχίας που υπερβαίνει το 72%, το οποίο είναι κατά πολύ υψηλότερο από το 20% της τυχαίας πρόβλεψης. Τα υπόλοιπα στατιστικά μεγέθη που συνοδεύουν το συγκεκριμένο πείραμα είναι ο σταθμισμένος μέσος όρος της ακρίβειας (precision) στα 0,74, της ανάκλησης (recall) στα 0,72 και της μέτρησης F (F-measure) στα 0,73. Σε ότι αφορά την περιοχή κάτω από την καμπύλη ROC (ROC Area), ο σταθμισμένος μέρος όρος της ανέρχεται σε 0,83. Τέλος, η τιμή του μέσου απόλυτου σφάλματος (mean absolute error) είναι 0,1113 και ο συντελεστής κάπα του Cohen είναι 0,6372.

Στα συμπεράσματα αυτής της έρευνας συμπεριλαμβάνεται και η ευθυγράμμιση των προδιαγραφών λειτουργίας του ταξινομητή με την πειραματική συμπεριφορά του. Δηλαδή, όπως αναμενόταν, το νευρωνικό δίκτυο με συνάρτηση ακτινωτής βάσης κατηγοριοποίησε τους χρήστες σύμφωνα με την ημερήσια χρήση υπολογιστή με υψηλό ποσοστό επιτυχίας και σε σύντομο χρονικό διάστημα.

5. ΣΥΝΟΨΗ – ΣΥΜΠΕΡΑΣΜΑΤΑ

Στην παρούσα εργασία η δυναμική της πληκτρολόγησης, δηλαδή η λεπτομερής χρονική καταγραφή των ενεργειών ενός χρήστη επί του πληκτρολογίου QWERTY, χρησιμοποιήθηκε για την κατηγοριοποίηση χρηστών. Με την κατηγοριοποίηση χρήστη, τη διαδικασία κατά την οποία ένας χρήστης υπολογιστή εντάσσεται σε μία κλάση, επιτυγχάνεται η πρόβλεψη κάποιων εγγενών ή επίκτητων χαρακτηριστικών του, όπως το φύλο, η ηλικία, η προτίμηση χεριού, το μορφωτικό επίπεδό του, κ.α. Για τον εντοπισμό των χαρακτηριστικών των χρηστών έχουν προταθεί και άλλες μέθοδοι, όπως η εξέταση των φωτογραφιών, των κειμένων και του κοινωνικού δικτύου τους, κάθε μία όμως από αυτές αντιμετωπίζει κάποιους περιορισμούς, με αποτέλεσμα την αδυναμία γενίκευσής της. Αντίθετα, με τη δυναμική της πληκτρολόγησης απορρέει μέθοδος που είναι ανεξάρτητη ομιλούμενης γλώσσας, δεν επεξεργάζεται ευαίσθητα ή και προσωπικά δεδομένα και απαιτεί μόνο τη παραγωγή του πιο συνηθισμένου μέσου επικοινωνίας μεταξύ των χρηστών, του κειμένου.

Το χαρακτηριστικό που εξετάστηκε ήταν η ημερήσια χρήση υπολογιστή, από το οποίο δεν προκύπτουν άμεσα συμπεράσματα για τα χαρακτηριστικά του χρήστη, αλλά χρήσιμες ενδείξεις για το φύλο, την ηλικία και το ετήσιο εισόδημά του, ενώ ταυτόχρονα παρέχονται και πληροφορίες εγκληματολογικού ενδιαφέροντος, αφού για παράδειγμα, η εγκληματολογική έρευνα θα απέκλειε όλους εκείνους τους υπόπτους που το πλήθος των ωρών ενασχόλησής τους με υπολογιστή ανά ημέρα, θα τους κατέτασσε στους μη εξειδικευμένους χρήστες και άρα όχι ικανούς για μια κακόβουλη ηλεκτρονική επίθεση. Η απόκτηση τέτοιου είδους πληροφοριών, σε συνδυασμό με άλλες που προκύπτουν και αυτές από την κατηγοριοποίηση χρηστών διά μέσου δυναμικής της πληκτρολόγησης, όπως για παράδειγμα το φύλο, η ηλικία, το μορφωτικό επίπεδο, το επικρατές χέρι και άλλα χαρακτηριστικά του ατόμου που διέπραξε μια κακόβουλη ενέργεια, θα καθιστούσαν την προτεινόμενη μέθοδο πολύτιμο εργαλείο στα χέρια των ειδικών της ψηφιακής εγκληματολογίας.

Το σύστημα που παρουσιάζεται εκμεταλλεύεται 60 διάρκειες πατήματος πλήκτρου και 140 λανθάνοντες χρόνους διγράμματος, ενώ υλοποιείται με τη βοήθεια ενός νευρωνικού δικτύου με συνάρτηση ακτινωτής βάσης. Τα αποτελέσματα έδειξαν ένα ποσοστό ορθής πρόβλεψης που ανέρχεται έως και το 75%, που είναι εμφανώς υψηλότερο από το 20% της τυχαίας πρόβλεψης.

Η πρωτοτυπία της εργασίας έγκειται στην εκμετάλλευση των παραμέτρων της δυναμικής της πληκτρολόγησης στην κατηγοριοποίηση χρηστών, καθώς επίσης και στην πρώτη απόπειρα εύρεσης της ημερήσιας χρήσης υπολογιστή. Σημαντικό συμπέρασμα αποτελεί επίσης το ότι οι διάρκειες πατήματος πλήκτρου και οι λανθάνοντες χρόνοι διγράμματος μπορούν να χρησιμοποιηθούν για την εύρεση κάποιων χαρακτηριστικών ενός άγνωστου χρήστη.

ABSTRACT

Internet is a tool where billions of people are now connected for purposes of communication, entertainment, work and education. But alongside the benefits, many dangers are lurking that may harm the users financially, morally and socially. Many of these risks come from malicious users and one of their “weapons” is that they can retain their anonymity. In this paper we propose a method that exploits the data generated by the way people type. The reason is to find some characteristics of users, so that to remove the complete anonymity and to be able to alert the unsuspecting users for the probability that their interlocutors misrepresent their characteristics. The proposed system was implemented with artificial neural networks and correct prediction exceeded 75%, well above than 20% of random prediction.

ΑΝΑΦΟΡΕΣ

Alrashed, H. F., & Berbar, M. A. (2013). Facial gender recognition using eyes images. *International Journal of Advanced Research in Computer and Communication Engineering, 2*(6), 2441-2445.

Bartlow, N., & Cukic, B. (2006). Evaluating the reliability of credential hardening through keystroke dynamics. *In Proceedings of 17th International Symposium on Software Reliability Engineering*. doi:10.1109/issre.2006.25.

Broomhead, D., & Lowe, D. (1988). Multivariable functional interpolation and adaptive networks. *Complex Systems, 2*, 321–355.

Cheng, N., Chandramouli, R., & Subbalakshmi, K. (2011). Author gender identification from text. *Digital Investigation, 8*(1), 78-88. doi:10.1016/j.diin.2011.04.002.

Choobeh, A. K. (2012). Improving automatic age estimation algorithms using an efficient ensemble technique. *International Journal of Machine Learning and Computing 2*(2), 118-122. doi:10.7763/ijmlc.2012.v2.99.

Daily computer usage in Great Britain by age 2006-2015. (n.d.). Retrieved from http://www.statista.com/statistics/275996/daily-computer-usage-penetration--in-great-britain-by-age/ (accessed on 15/04/2016).

Damayanti, F., & Rachmad, A. (2016). Recognizing gender through facial image using support vector machine. *Journal of Theoretical and Applied Information Technology, 88*(3), 607-612.

Douhou, S., & Magnus, J. R. (2009). The reliability of user authentication through keystroke dynamics. *Statistica Neerlandica, 63*(4), 432-449. doi:10.1111/j.1467-9574.2009.00434.x.

Hassan, S., Selim, M., & Zayed, H. (2013). User authentication with adaptive keystroke dynamics. *International Journal of Computer Science Issues, 10*(4), 127-134.

Hewahi, N., Olwan, A., Tubeel, N., EL-Asar, S., & Abu-Sultan, Z. (2010). Age estimation based on neural networks using face features. *Journal of Emerging Trends in Computing and Information Sciences, 1*(2), pp. 61-67.

Hofferth, S., Flood, S., & Sobek, M. (2013). American time use survey data extract system: Version 2.4 [Machine-readable database]. Maryland Population Research Center, University of Maryland, College Park, Maryland, and Minnesota Population Center, University of Minnesota, Minneapolis, Minnesota.

Hosseinzadeh, D., & Krishnan, S. (2008). Gaussian mixture modeling of keystroke patterns for biometric applications. *IEEE Transactions on Systems, Man, and Cybernetics, 38*(6), 816-826. doi:10.1109/tsmcc.2008. 2001696.

Internet live stats. (n.d.). Retrieved from http://www.internetlivestats.com (accessed on 02/02/2017).

Kumar, A., Patwari, A., & Sabale, S. (2014). User authentication by typing pattern for computer and computer based devices. *International Journal of Advanced Research in Computer and Communication Engineering, 3*(10), 8132-8134. doi:10.17148/ijarcce.2014.31011.

Lachenbruch, P. A., & Mickey, M. R. (1968). Estimation of error rates in discriminant analysis. *Technometrics, 10*(1), 1. doi:10.2307/1266219.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *In Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, 1*, 281-297.

Mukherjee, A., & Liu, B. (2010). Improving gender classification of blog authors. *In Proceedings of Conference on Empirical Methods in Natural Language Processing*, 207-217.

Patil, R., & Renke, A. (2016). Keystroke dynamics for user authentication and identification by using typing rhythm. *International Journal of Computer Applications, 144*(9), 27-33. doi:10.5120/ijca2016910432.

Rao, D., Yarowsky, D., Shreevats, A., & Gupta, M. (2010). Classifying latent user attributes in twitter. *In Proceedings of 2nd International Workshop on Search and Mining User-Generated Contents*, 37-44. doi:10.1145/1871985.1871993.

Schler, J., Koppel, M., Argamon, S., & Pennebaker, J. (2006). Effects of age and gender on blogging. *In Proceedings of AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*, 199-205.

Suler, J. (2004). The online disinhibition effect. *CyberPsychology & Behavior, 7*(3), 321-326. doi:10.1089/1094931041291295.

Wankhede, S., & Verma, S. (2014). Keystroke dynamics authentication system using neural network. *International Journal of Innovative Research and Development, 3*(1), 157-164.

Winn, J., & Heeter, C. (2009). Gaming, gender, and time: Who makes time to play? *Sex Roles, 61*(1-2), 1-13. doi:10.1007/s11199-009-9595-7.

Yi, D., Lei, Z., & Li, S. Z. (2014). Age estimation by multi-scale convolutional network. *In Proceedings of 12th Asian Conference of Computer Vision,* 144-158. doi:10.1007/978-3-319-16811-1_10.

Zhao, Y. (2006). Learning user keystroke patterns for authentication. *In Proceedings of World Academy of Science, Engineering and Technology, 14*, 65-70.