УДК 004.93+510

ІДЕНТИФІКАЦІЯ КОРИСТУВАЧІВ КОРПОРАТИВНОЇ МЕРЕЖІ З ВИКОРИСТАННЯМ ЛІНГВІСТИЧНОГО МОДЕЛЮВАННЯ

Василенко В.Г., Ширій В.В.
Національний технічний університет України
«Київський політехнічний інститут»

Розглядається використання лінгвістичного моделювання, як одного з напрямків нечисельного моделювання, для ідентифікації користувача у корпоративній мережі. Описується загальна структура системи ідентифікації та алгоритм реалізації методу на базі інтервального підходу, в основі якого лежить процес відновлення формальної граматики.
Ключові слова: Лінгвістичне моделювання, біометрична ідентифікація, розпізнавання образів, інтервальний підхід.

Постановка проблеми. Комп'ютери та послуги, які ідентифікують користувачів тільки при вході в систему за допомогою облікових даних, уразливі для крадіжки особистих даних. Хакери можуть здійснювати шахрайську діяльність під викраденими захисними даними, таких як паролі чи смарткарт, які буди отримані незаконно або за допомогою комп'ютерів, які залишаються без нагляду. Призначений для користувача метод перевірки забезпечує додатковий шар безпеки, підтверджуючи ідентичність зареєстрованих користувачів. Ми використаємо метод, який перевіряє користувачів відповідно до особливостей їх взаємодії з маніпулятором комп'ютера під назвою миша.

Аналіз останніх досліджень і публікацій. Найбільш поширені поведінкові біометричні методи верифікації засновані за:

– допомогою руху миші [1; 2; 3; 4], які випливають із взаємодії користувача миші;

– динамікою натискань клавіш [6; 7; 8], які отримані від активності натискання користувачем клавіатури;

– взаємодією програмного забезпечення, які покладаються на особливості, взяті із взаємодії користувача з конкретним програмним засобом.

Виділення не вирішених раніше частин загальної проблеми. Опис способу використання лінгвістичного моделювання для ідентифікації користувачів, що під'єднані до корпоративної мережі. Наведено схему поведінкової системи біометричної ідентифікації та алгоритму моделювання.

Мета статті. Головною метою цієї роботи є використання лінгвістичного моделювання для ідентифікації користувача за допомогою руху миші.

Виклад основного матеріалу. В даний час більшість комп'ютерних систем і веб-сайтів онлайн, ідентифікує користувачів за допомогою логіну і паролю. Але зазвичай, хакери можуть легко вкрасти пароль за допомогою великої кількості методів. Деякі з цих методи – напади фішингу, кейлогерів тощо. Також комп'ютер користувача може залишатися невимкнутим в той час, коли хакери зможуть встановити кейлогер на комп'ютер або надіславши деякі посилання на цей комп'ютер, наприклад, привітання або зображення і тд. Якщо користувач натискає на ці посилання, кейлогер починає записувати кожне натискання клавіш, що в подальшому можуть містити логін та пароль користувача, та скріншоти після декількох хвилин роботи. Також вони можуть посилати дані хакерам, не повідомляючи про це користувача.

Згідно некомерційної організації Центр Ідентифікації Крадіжок (Identity Theft Resource Center, ITRC), так звана «крадіжка особистості», з споживчої точки зору, розділяється на чотири категорії:

– Фінансова «крадіжка особи», в якій вкрадена особистість використовується, щоб отримати товари і послуги;

– Злочинна «крадіжка особи», в якій злочинець виконують роль законного користувача, коли станеться злочин;

– Клонування ідентичності – використання інформації іншої людини, щоб прийняти його або її особистість в повсякденному житті;

– Ділова/комерційна «крадіжка особи» – використання вкраденого бізнес-імені, щоб отримати кредит.

Недолік ідентифікаційних методів, які засновані лише на основі облікових даних призводять до введення користувальницьких методів ідентифікації і перевірки. Засновані вони на поведінковій і фізіологічній біометрії, які, як передбачається, унікальні один від одного і саме через це, в майбутньому, дані важко вкрасти. Автентифікація виконується лише один раз на початку сеансу, в той час як, перевірка ідентичності виконується безперервно протягом сесії. Перевірка ідентичності може бути досягнута за допомогою одного з двох методів: поведінкової або фізіологічної біометричної системи. Поведінкова біометрична особливість включає особливості взаємодії користувача і пристроїв введення, таких як миша і клавіатура. А фізіологічне використовує людські особливості, які є унікальними для індивіда. Для прикладів: відбитки пальців, візерунки райдужної оболонки ока, обличчя, рухи губ, хода/крок, голос/мова, підпис/почерк тощо.

Таким чином системи, що використовують біометричну перевірку користувача, вимагають хакера, який зможе не тільки просочитися в систему, щоб вкрасти облікові дані користувача, але також і наслідувати користувацьку поведінкову або фізіологічну біометрію, що робить викрадення ідентифікаційних даних набагато важчою задачею. Ми зосереджені на поведінковій системі перевірки, тому що вона, на відміну від фізіологічної перевірки, не вимагає додаткових апа-

ратних засобів. Очевидно і те, що вартість такої системи не може бути більшою.

Розглянемо загальну модель методу перевірки, яка заснована на русі миші користувача. Цей метод вимагає збереження і обробки десятків координаті дій миші, перш ніж можливо виконати перевірку. Перевірка кожної окремої дії миші підвищує точність при одночасному скороченні часу, який необхідний для перевірки автентичності користувача. В порівнянні з підходом, який показаний на гістограмі в [1], менше дій потрібно для досягнення певного рівня точності.

Загальну блок-схему запропонованої системи показано на рис. 1.

Рис. 1 зображує типову архітектуру поведінкової користувальницької системи перевірки біометрії. Система включає такі компоненти:

– Отримання подій – оброблює події, вироблені пристроєм введення. В нашому випадку, це є мишка. Події можуть бути рухом миші (Mouse Move, MM), рух вниз ліво (left down, LD), вліво вгору (left up, LU), вправо вниз (right down, RD), вправо вверх (right up, RU), очікування руху (silence, S) і тд.

– Виділення ознак – будує підпис, який характеризує поведінкові біометричні дані користувача. Ознаки можуть включати послідовність переміщення миші (MMS), натискання лівої (LC) та правої кнопки миші (RC) тощо.

– Класифікатор – складається з індуктора (наприклад, Векторні Машини Підтримки, Штучні нейронні мережі, тощо), який використовується для побудови моделі перевірки користувача для перевірки підписів. Під час перевірки, модель використовується для класифікації нових зразків підписів, отриманих від користувача.

– База даних Підписів – база даних поведінкових підписів, заснованих на навчанні моделі. Якщо кілька користувачів використовують комп'ютер, після введення логіну одного з користувачів, буде відновлено процес перевірки підписів.

В Базі Даних, підпис буде складатися з:

– кількості рухів мишкою;

– кількості натискання лівої та правої кнопки;

– часових інтервалів, коли мишка не рухається;

– агрегування координат миші.

Деякі типи підписів будуть створені для кожної сесії.

Біометрична система перевірки користувача – система розпізнавання образів, яка отримує біометричні дані від людини, визначає набір ознак, і встановлює унікальний, призначену для користувача, підпис і будує модель перевірки, щоб класифікувати підписи між різними користувачами. У вищезгаданому рис. 1:

– Зелений сигнал – Зареєстрований користувач;

– Червоний сигнал – Неавторизований користувач, можливо зловмисник.

Основним завданням лінгвістичного моделювання є перетворення чисельних рядів, масивів координат до лінгвістичних послідовностей та відновлення за ним формальної граматики мови відповідного характеру для вирішення проблем які виникли: аналіз та прогнозування часового ряду, автентифікація користувача за його рухами [10].

Лінгвістичне моделювання базується на трьох основних підходах: структурний підхід та математична лінгвістика, інтервальні обчислення та робастні методи, сучасні методи ймовірнісного моделювання.

В основі лінгвістичного моделювання лежить лема існування ізоморфізма відтворення чисельних даних до лінгвістичних послідовностей, на основі яких може бути побудована мова. Як висновок існування унікальної мови, яка фактично уособлюється наборами чисельних даних.

Початок цього був покладений при створенні математичних основ автентифікації користувача складною технічною системою. Автентифікація – шлях встановлення вірогідності інформації, пред'явленої користувачем у разі звернення його до системи та відкриття йому доступу, якщо він має на це право. Дамо загальну постановку завдання. Аналізується кінематика рухів користувача при керуванні складною технічною системою. В нашому випадку – це оператор, що сидить за комп'ютером та взаємодіє з корпоративною мережею. При цьому стоїть необхідність автентифікації користувача – в цей момент керує складною технічною системою оператор А чи не А.

Вхідними даними для моделювання є данні зняті з руху миші довжиною M: $X = \{X_i\}_{i=1}^{M} = \{X_1, X_2, \ldots, X_i, \ldots, X_M\}$, який описує деякий динамічний процес.

Для початку необхідно побудувати на основі часового ряду $X = \{X_i\}_{i=1}^{M}$ різницевих рядів X_1, X_2, ...:

$$X_i^1 = X_{i+1} - X_i$$
$$X_i^2 = X_{i+1}^1 - X_i^1$$
....

Після чого сортуємо ряд за зростанням $X^1 \rightarrow X^{s1}$ та знаходимо max (X^{s1}) та min (X^{s1}). Розбиваємо відрізки [min (X^{s1}),0] та [0,max (X^{s1})] на N відрізків за правилами інтервалізації, відповідно до свого варіанту. N змінюється від 10 до

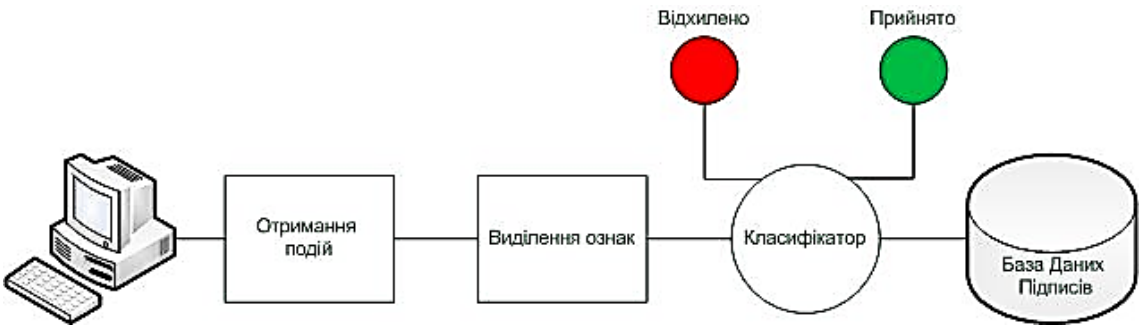

Рис. 1. Типова структура поведінкової системи біометричної ідентифікації

Джерело: [9]

33 цей параметр залежності від алфавіту, який буде обраний на етапі лінгвістизації.

Розбиття на інтервали відбувається таким чином, щоб кількість елементів різнецевого ряду в кожний інтервал потрапляла у відповідності до певного розподілу. Тобто частота попадання елементів до інтервалу [a,b] дорівнювала теоретичній ймовірності $P\{x\in [a,b]\} = F(b) - F(a)$, де F-функція відповідного розподілу.

В результаті отримуємо дві множини інтервалів:

$I_{0,1} = [a_0,a_1]$, $I_{1,2} = [a_1,a_2]$, ..., $I_{N-2,\ N-1} = [a_{N-2},a_{N-1}]$, $I_{N-1,\ N} = [a_{N-1},a_N]$,

де $a_0 = \min(X^{s1})$, $a_N = 0$;

$J_{0,1} = [b_0,b_1]$, $J_{1,2} = [b_1,b_2]$, ..., $J_{N-2,N-1} = [b_{N-2},b_{N-1}]$, $J_{N-1,N} = [b_{N-1},b_N]$,

де $b_0 = 0, b_N = \max(X^{s1})$;

В подальшому необхідно відсортуємо символи алфавіту у наступному порядку: $a_1 = z$, $a_2 = y$, ..., $a_{N-1} = b$, $a_N = z$, $a_{N+1} = A$, $a_{N+2} = B$, ..., $a_{2N-1} = Y$, $a_{2N} = Z$.

Далі будується відображення L: $X^1 \rightarrow Y$ за такими правилами:

$$L(x_i) = \begin{Bmatrix} a_j \text{ якщо } x_i \epsilon I_{j-1,j} \\ a_{N+j} \text{ якщо } x_i \epsilon J_{j-1,j} \end{Bmatrix}$$

Застосувавши відображення L, до елементів ряду X1. В результаті чого отримуємо ряд $L(x_1^1), \ldots, L(x_M^1)$.

Будуємо матрицю передування для прихованої марковської моделі. Множина станів – це обраний нами алфавіт. Для кожної пари станів, наприклад $<d,S>$ підраховуємо скільки разів вона зустрічається в лінгвістичному ланцюжку $L(x_1^1)\ldots L(x_M^1)$. Поділивши $V_{d,S}$ на загальну кількість входжень літери «d» дотримуємо частоту переходів зі стану «d» в стан «S»: $v(d \rightarrow S) = \frac{v_{d,S}}{w_d}$.

Далі знаходимо в лінгвістичному ланцюгу повтори переходу від двох, трьох та більше станів. Далі необхідно побудувати розширену матрицю, додавши до станів варіанти двох, трьох та більше станів, що зустрічаються в нашому лінгвістичному ланцюгу. Після чого побудувати по розширеній матриці передування правила ймовірнісної граматики. Тобто для кожної ненульової клітинки будується правило наступного вигляду: $Sax \overset{0.5}{\rightarrow} Z$.

Алфавіт та правила передування утворюють лінгвістичну модель ряду X^1. Ту саму процедуру побудови лінгвістичної моделі повторюємо для інших різниць ряду X – X^2, X^3, X^4, X^5, X^6. Будуємо лінгвістичну модель для алфавіту потужності – 10, 15, 20, 26. Та аналізуємо відмінностей результатів лінгвістичного моделювання одного й того ж самого чисельного ряду, які виникають при двох різних правилах інтервалізації.

Висновки і пропозиції. Було розглянуто загальну модель перевірки біометричної ідентифікації користувача, наведено компоненти цієї системи. Кожна з цих компонент виконує свою певну функцію. Наприклад, База Даних Підписів зберігає поведінкові підписи моделі, що дозволяє виявляти сторонніх користувачів чи користувачів, що зареєстровані, однак використовують інший обліковий запис.

Лінгвістичне моделювання використовуючи перетворення чисельних рядів чи багатовимірних даних в лінгвістичні послідовності, перетворює ці послідовності в формальну граматику. За допомогою цього моделювання, можливо ідентифікувати користувача в корпоративній мережі та виявляти зловмисників, які могли вкрасти дані працівників компанії.

Однак, лінгвістичне моделювання можливо використовувати і для визначення емоційного стану користувача, діагностування хвороб, зв'язаних з рухом кінцівок (тремор, тік чи хорея).

Список літератури:

1. Clint Feher, Yuval Elovici, Robert Moskovitch, Lior Rokach, Alon Schclar, «User identity verification via mouse dynamics», Information Sciences 201 (2012) 19–36.
2. Chao Shen, Zhongmin Cai, Xiaohong Guan, Youtian Du, and Roy A. Maxion, User Authentication Through Mouse Dynamics. IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITYVOL. 8, № 1, Jan 2013.
3. Zach Jorgensen and Ting Yu, On Mouse Dynamics as a Behavioral Biometric for Authentication. ACM 978-1-4503-0564-8/March 2011.
4. Jorgensen Z., Yu T. «On mouse dynamics as a behavioral biometric for authentication, in: Proceedings of the Sixth ACM Symposium on Information, Computer, and Communications Security» (AsiaCCS), March 2011.
5. Saurabh Singh, Dr. K. V. Arya, «Mouse Interaction based Authentication System by Classifying the Distance Traveled by the Mouse» International Journal of Computer Applications (0975–8887) Volume 17. – № 1, March 2011.
6. Lívia C. F. Araújo, Luiz H. R. Sucupira Jr., Miguel G. Lizárraga, Lee L. Ling and Joro B. T. Yabu-Uti, «User Authentication through Typing Biometrics Features, IEEE Transactions on Signal Processing», Vol. 53, № 2, February 2005.
7. Cho S., Han C., Han D. H., Kim H.I. «Web-based keystroke dynamics identity verification using neural network, Journal of Organizational Computing and Electronic Commerce» 10 (4) (2000) 295–307.
8. Ballard L., Lopresti D., Monrose F. «Evaluating the security of handwriting biometrics, in: The 10th International Workshop on Frontiers in Handwriting Recognition» (IWFHR 06), La Baule, France, 2006.
9. Gorad B. J., Kodavade D. V. «User Identity Using Mouse Signature, IOSR Journal of Computer Engineering», Volume 12, Issue 4, Jul.–Aug. 2013, Pp. 33-36.
10. Баклан І. В. Лінгвістичне моделювання: основи, методи, деякі прикладні аспекти / І. В. Баклан // Систем. технології. – 2011. – № 3. – С. 10-19. – Бібліогр.: 9 назв. – укр.
11. Баклан І. В., Петренко О. О., Селін Ю. М. Структурний підхід до розпізнавання образів у системах безпеки / Національна безпека України: стан, кризові явища, шляхи її подолання. – 2005. – С. 375-380.
12. Баклан І. В. Використання ймовірнісних моделей для аутентифікації оператора складної технічної системи / Національна безпека України: стан, кризові явища, шляхи її подолання. – 2005. – С. 380-386.
13. Баклан І. В. Лінгвістичне моделювання числових рядів різної природи з фрактальними властивостями / Системні технології. – 2016 – С. 110-118.

Василенко В.Г., Ширий В.В.
Национальный технический университет
«Киевский политехнический институт»

ИДЕНТИФИКАЦИЯ ПОЛЬЗОВАТЕЛЕЙ КОРПОРАТИВНОЙ СЕТИ С ИСПОЛЬЗОВАНИЕМ ЛИНГВИСТИЧЕСКОГО МОДЕЛИРОВАНИЯ

Аннотация
Рассматривается использование лингвистического моделирования, как одного из направлений численного моделирования, для идентификации пользователя в корпоративной сети. Описывается общая структура системы идентификации и алгоритм реализации метода на базе интервального подхода, в основе которого лежит процесс восстановления формальной грамматики.
Ключевые слова: Лингвистическое моделирование, биометрическая идентификация, распознавание образов, интервальный подход.

Vasilenko V.G., Shyrii V.V.
National Technical University of Ukraine
«Kyiv Polytechnic Institute»

IDENTIFICATION OF USERS OF A CORPORATE NETWORK WITH USE OF LINGUISTIC MODELING

Summary
Examines the use of linguistic modeling, as one of the areas of numerical modeling, to identify the user on the corporate network. Describes the general structure of the system identification algorithm and the realization method based on interval approach, which is based on the recovery process of a formal grammar.
Keywords: Linguistic modeling, biometric identification, pattern recognition, interval approach.