UserPrint: Perfilização e Distinção de Comportamento Usuários em Ambientes Corporativos

Barbara Reis dos Santos
Universidade Federal do Paraná – UFPR
Curitiba, PR, Brasil
brs22@inf.ufpr.br

Gabriela Fanaia de Almeida Dias Dorst
Universidade Federal do Paraná – UFPR
Curitiba, PR, Brasil
gfadd22@inf.ufpr.br

Simone Dominico
Universidade Federal do Paraná – UFPR
Curitiba, PR, Brasil
simone@inf.ufpr.br

André Ricardo Abed Grégio
Universidade Federal do Paraná – UFPR
Curitiba, PR, Brasil
gregio@ufpr.br

Paulo Lisboa de Almeida
Universidade Federal do Paraná – UFPR
Curitiba, PR, Brasil
paulorla@ufpr.br

Abstract

This article introduces *UserPrint*, a behavioral fingerprinting approach that characterizes corporate users based on their software usage patterns. A real dataset comprising 41,859 users collected over six months is analyzed to evaluate the temporal consistency of application usage through cosine similarity and cosine distance. The representation is further extended with web navigation signals using a controlled synthetic dataset derived from a real user profile, enabling the assessment of how additional modalities affect profile separability. The results demonstrate that cosine-based metrics reliably capture both stable behaviors and meaningful temporal changes, while the inclusion of navigation data substantially increases user separability. These findings indicate that continuous behavioral metrics and multimodal feature representations provide a promising foundation for security monitoring and risk detection in corporate environments.

Keywords

Perfilização comportamental, Detecção de intrusão, Análise de comportamento de usuários, Similaridade do cosseno, Segurança corporativa

1 Introdução

O comportamento dos usuários no ambiente digital corporativo é uma variável crítica para a segurança da informação das empresas. A forma como indivíduos interagem com softwares ao longo do tempo pode revelar não apenas padrões operacionais, mas também mudanças que indiquem riscos, violações de política ou até mesmo ações maliciosas. Nesse contexto, a análise temporal do uso de aplicações torna-se uma ferramenta estratégica para a detecção de anomalias e a antecipação de ameaças internas, com a modelagem de perfis de comportamento ao longo do tempo.

Logo, a identificação precisa de usuários com base em seus padrões de uso de aplicações é importante para mitigar riscos de segurança, como acessos não autorizados, *insider threats* e comprometimento de credenciais. Métodos tradicionais de autenticação (como senhas e tokens) são estáticos e vulneráveis a violações, enquanto a análise comportamental contínua pode fornecer uma camada adicional de segurança adaptativa. Além disso, a escalabilidade da detecção em ambientes corporativos com milhares de usuários exige abordagens automatizadas e eficientes, capazes de lidar com grandes volumes de dados sem sacrificar a precisão.

Com o avanço das soluções de monitoramento em *endpoint*, tornou-se viável coletar dados detalhados sobre o uso de aplicações e URLs (*Uniform Resource Locator*, ou endereços web) em escala organizacional. No entanto, a extração de informações relevantes desses dados ainda representa um desafio, especialmente na avaliação da consistência do comportamento ao longo do tempo.

Além disso, apesar dos avanços em monitoramento de *endpoint* e análise comportamental, a literatura ainda apresenta lacunas significativas. Primeiro, muitas abordagens dependem de regras pré-definidas ou modelos estáticos, que não capturam a evolução natural do comportamento dos usuários. Segundo, técnicas baseadas em agregação temporal (como médias diárias ou semanais) podem ocultar variações sutis, porém críticas, no uso de aplicações. Terceiro, há uma carência de estudos empíricos em larga escala que validem métricas de similaridade em cenários reais e dinâmicos, especialmente considerando a diversidade de perfis de usuários em ambientes corporativos.

Neste artigo, apresenta-se o *UserPrint*, um estudo empírico sobre a evolução do comportamento de usuários com base na vetorização de padrões de uso de aplicações, comparado ao longo do tempo com similaridade do cosseno e distância do cosseno. Foram utilizados dados de 41.859 usuários ao longo de seis meses, contendo informações sobre aplicações. A avaliação foi complementada com sinais de navegação web por meio de dados sintéticos gerados a partir de um único perfil real, para uma avaliação controlada, discutindo as limitações dessa escolha. Avaliou-se a eficácia das métricas utilizadas para comparar períodos de atividade. O objetivo consiste em responder às seguintes perguntas de pesquisa:

(RQ1) Quão estáveis são os perfis de uso de aplicações ao longo do tempo em escala corporativa?

(RQ2) Como a inclusão de sinais de navegação afeta a separabilidade/estrutura de vizinhança entre perfis?

Os resultados indicam que a similaridade do cosseno captura padrões consistentes e mudanças significativas no comportamento.

Utilizando a distância do cosseno foi possível evidenciar variações mais sutis. Os resultados evidenciam que a aplicação conjunta dessas medidas, capturam padrões relevantes de mudanças relevantes, sendo adequadas para cenários de segurança comportamental e predição de riscos.

As principais contribuições deste artigo incluem, a avaliação em dados corporativos reais, da consistência de perfis de uso de aplicações por meio da similaridade e distância do cosseno, assim como a demonstração de que essas medidas são capazes de capturar tanto estabilidade quanto mudanças de comportamento. Esses resultados evidenciam o potencial dessas abordagens como componentes de apoio a processos de monitoramento e detecção de anomalias em ambientes corporativos.

Este artigo está organizado da seguinte forma: a Seção 2 apresenta os trabalhos relacionados; a Seção 3 descreve o protocolo experimental e os dados utilizados; a Seção 4 detalha os conceitos fundamentais; a Seção 5 aborda os experimentos e métricas aplicadas; e a Seção 6 apresenta a conclusão e trabalhos futuros.

2 Trabalhos Relacionados

A literatura sobre autenticação contínua (CA) e análise comportamental é vasta, abrangendo diversas fontes de dados como tráfego de rede [1], dinâmicas de digitação [2] e uso de aplicações [3]. No entanto, como destacado na introdução, persistem lacunas significativas que limitam a aplicabilidade prática desses estudos. Entre elas, destacam-se: (i) a escassez de *datasets* públicos diversificados coletados em ambientes reais; (ii)a dependência de modelos estáticos que não capturam a evolução natural do comportamento dos usuários [4, 5].

Estudos recentes ilustram essas limitações. Dave et al. investigaram a autenticação baseada na movimentação do mouse, porém com apenas 19 usuários em cenários restritivos de jogos. De forma similar, Mondal and Bours coletaram dados de teclado e mouse em ambiente não controlado, mas a extração de características por ação inviabiliza sua aplicação em sistemas reais devido ao *overhead* computacional. Esses trabalhos muitas vezes negligenciam a robustez temporal e a escalabilidade.

Para superar tais limitações, a área de *User and Entity Behavior Analytics* (UEBA) emerge como uma abordagem promissora, focando na identificação de anomalias através da análise de padrões comportamentais [8]. No entanto, técnicas comuns em UEBA, como *clustering* [9], são sensíveis a ruídos e frequentemente geram falsos positivos quando confrontadas com variações comportamentais legítimas (*concept drift*).

A escolha da métrica é um fator crítico. Um aspecto relacionado às lacunas de modelagem temporal e escalabilidade, é a seleção de métricas adequadas para comparação de perfis comportamentais. Enquanto trabalhos anteriores frequentemente adotam métricas binárias como o Índice de Jaccard [10], que considera apenas a presença ou ausência de aplicações, tais abordagens ignoram a informação sobre a intensidade do uso, limitando sua sensibilidade para detectar mudanças sutis.

Métricas contínuas oferecem uma alternativa mais rica. A Distância Euclidiana, embora intuitiva, é sensível à magnitude dos dados. Como demonstrado por Aggarwal, em espaços de alta dimensionalidade, essa sensibilidade pode gerar falsos positivos. Um pico legítimo de atividade (ex.: fechamento de trimestre) aumenta a magnitude do vetor de uso, fazendo com que o usuário seja erroneamente sinalizado como anomalia, embora seu *padrão relativo* de uso permaneça inalterado.

Por outro lado, a similaridade do cosseno captura a forma do perfil, sendo invariante a variações de escala. Seu uso em Análise de Linguagem Natural [12] fornece uma direção, isso porque, assim como vetores de palavras são comparados pelo contexto em que aparecem, perfis de uso de aplicações podem ser comparados pelo contexto funcional que representam. Dois usuários com magnitudes muito diferentes de uso, mas proporções semelhantes entre aplicações, permanecem próximos nessa métrica, o que é desejável em cenários de caracterização de função ou papel.

A Distância do Cosseno, derivada diretamente da Similaridade do Cosseno, preserva a mesma noção de proximidade angular, mas em uma forma adequada a métodos que requerem uma função de distância, como *DBSCAN* e *k-NN* [13]. Por isso, é utilizada em tarefas de agrupamento e detecção de anomalias, permitindo aplicar a comparação angular entre perfis em modelos baseados em vizinhança.

Reconhecendo os pontos fortes e fracos de cada abordagem, este trabalho não as vê como excludentes, mas como complementares. Investigou-se a hipótese de que a Similaridade do Cosseno e a Distância do Cosseno serve como a espinha dorsal para a UserPrint, definindo a identidade comportamental fundamental do usuário de forma robusta. Esta abordagem híbrida busca mitigar as limitações de cada métrica isoladamente, oferecendo um modelo mais preciso e com menor taxa de falsos positivos para segurança comportamental.

3 Protocolo Experimental

Esta seção apresenta o protocolo experimental, com o detalhamento da base de dados e as estratégias de representação utilizadas.

3.1 Coleta e Armazenamento de Dados

Neste trabalho foram utilizadas duas representações distintas de comportamento de usuários. A primeira, denominada **Base Simples**, contém exclusivamente dados de uso de aplicações, incluindo nome e versão do aplicativo, duração das sessões e identificadores anonimizados de usuários e empresas. A segunda, denominada **Base Composta**, estende esse conjunto ao incorporar informações de navegação web, adicionando URLs completas, domínios e aplicações associadas a requisições HTTP e HTTPS capturadas pelo módulo de *web filter*.

Os dados brutos da Base Simples foram coletados por meio de um agente de monitoramento de segurança instalado, com consentimento, nos dispositivos de usuários corporativos. A coleta abrangeu o período de 21 de março a 30 de setembro de 2024, totalizando 407 GB de dados distribuídos em 2.784.524.189 registros de uso, referentes a 41.859 usuários de 709 empresas. No total, foram identificadas 43.857 aplicações distintas, das quais instaladores e programas auxiliares foram filtrados por não contribuírem para a caracterização comportamental. O armazenamento e processamento dos dados foram realizados no PostgreSQL 14.11.

A Base Composta utilizou dados provenientes de um arquivo JSON contendo 28.028 eventos de navegação web registrados em 11 de julho de 2025, entre 17:05:17 e 20:15:30. Cada evento inclui

informações como URL acessada, domínio, aplicação responsável pela requisição, carimbo de tempo e identificadores de usuário. Esse arquivo corresponde à navegação completa de um único usuário de uma única empresa, o qual é adotado neste trabalho como **usuário base**. No total, foram identificadas 23 aplicações distintas, 8.373 URLs únicas e 530 domínios.

Esse conjunto de navegação foi utilizado como referência para a geração de dados sintéticos simulando múltiplos usuários da mesma organização. Para isso, foram identificados, na Base Simples, os 10 usuários mais semelhantes ao usuário base segundo o padrão de uso das aplicações. A partir dessa seleção, foram sintetizados dados de navegação para esses mesmos usuários, que não possuíam registros na Base Composta. Esse procedimento preserva a coerência estrutural entre padrões de uso de aplicações e padrões de navegação, permitindo avaliar de forma controlada como a introdução de URLs altera a separabilidade entre perfis comportamentais.

3.2 Definição dos experimentos

O estudo foi estruturado em dois experimentos utilizados para validar a abordagem UserPrint em diferentes níveis de granularidade.

O **Experimento 1** teve como objetivo principal avaliar a generalização e robustez do método de *similaridade do cosseno* em larga escala. Para isso, utilizou-se o conjunto completo de dados reais de uso de aplicações, abrangendo usuários de todas as 709 empresas. O foco desta etapa foi a criação e validação de perfis individuais baseados exclusivamente no uso de aplicações, maximizando a quantidade de usuários analisados para testar a consistência metodológica em um cenário amplo e heterogêneo.

O **Experimento 2** concentrou-se em reproduzir o processo de comparação entre usuários em um ambiente corporativo específico usando *Distância de cosseno*, integrando informações de uso de aplicações e dados de navegação web. O ponto de partida foi o **usuário base**, cuja atividade registrada permite reconstruir seu perfil completo de uso de aplicações, servindo como referência para geração e comparação dos demais perfis.

Com o vetor de características do **usuário base** definido, foram identificados os dez usuários mais semelhantes por meio da distância do cosseno, considerando representações absoluta e proporcional do perfil. Como a Base Composta continha dados de navegação apenas para o usuário base, foram gerados perfis sintéticos para os dez usuários mais próximos, combinando seus padrões dominantes de uso de aplicações com padrões de acesso a URLs derivados do comportamento real do usuário base.

A Base Composta foi então representada vetorialmente com base em número de ocorrências e diversidade de domínios por aplicativo. Por fim, realizaram-se projeções em PCA e cálculos de proximidade para ambas as representações, possibilitando a comparação da estrutura de vizinhança entre a Base Simples e a Base Composta acrescida dos perfis sintéticos.

4 Conceitos Fundamentais e Metodologia

4.1 Perfilização Comportamental

A perfilização comportamental consiste na identificação e modelagem dos hábitos de um indivíduo com base em suas interações digitais. Em ambientes corporativos, essa técnica permite construir um "perfil de uso" que pode ser utilizado como referência para detectar desvios comportamentais. Essa abordagem é amplamente utilizada em sistemas de *User and Entity Behavior Analytics* (UEBA), com objetivo de monitorar comportamentos e prever riscos a partir do histórico de uso dos usuários.

4.2 Representações Absoluta e Proporcional

As representações **absoluta** e **proporcional** correspondem a duas formas distintas de modelar os vetores de características dos usuários, a principal diferença é o tratamento da magnitude dos dados.

Representação absoluta. Na representação absoluta, os vetores são construídos utilizando diretamente os valores reais observados nos dados. No caso da Base Simples, isso corresponde ao tempo total de uso de cada aplicação e à quantidade de eventos registrados. Já na Base Composta, os valores absolutos representam o número de ocorrências de cada aplicação e a quantidade total de domínios distintos acessados por meio de cada uma. Essa representação preserva integralmente a informação de volume de uso, sendo, portanto, sensível a diferenças de intensidade entre os usuários.

Representação proporcional. Na representação proporcional, os vetores são normalizados de forma que cada posição passe a representar a proporção relativa de cada aplicação no perfil do usuário. Nesse caso, os valores absolutos são divididos pela soma total dos respectivos vetores, resultando em uma distribuição que enfatiza o peso relativo de cada aplicação no comportamento do usuário, independentemente do volume total de uso. Essa abordagem reduz a influência de usuários com maior ou menor atividade global, permitindo comparações baseadas predominantemente no padrão de distribuição do uso.

Em síntese, a representação absoluta privilegia a **intensidade global de uso**, enquanto a proporcional enfatiza o **padrão relativo de distribuição**, tornando ambas complementares na análise desenvolvida neste trabalho.

4.3 Vetores de Características

No **Experimento 1**, o vetor de características foi construído a partir do tempo de uso das aplicações. Considerando uma lista fixa de n aplicações (por exemplo, 50), cada vetor terá n posições, onde cada posição corresponde a uma aplicação específica. Se um usuário utilizou a aplicação A por 300 segundos, a B por 600 segundos, e não utilizou as demais, o vetor correspondente será:

$$\mathbf{v} = [300,\ 600,\ 0,\ 0,\ \ldots,\ 0]$$

A partir desses vetores diários, foi calculado um vetor médio para o primeiro período (janela de treino), refletindo o padrão típico de uso do usuário. Em seguida, os vetores diários do segundo período (janela de teste) foram utilizados para comparar a consistência do comportamento ao longo do tempo, por meio da métrica de Similaridade do Cosseno. Essa abordagem permite avaliar a estabilidade do comportamento de uso ao longo do tempo, comparando dias posteriores com o padrão médio previamente estabelecido.

A Figura 1 ilustra esse processo de forma esquemática: os vetores diários do bloco de treino são utilizados para gerar um vetor médio, que é posteriormente comparado individualmente com cada vetor diário da janela de teste.

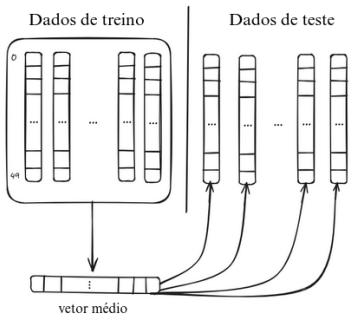

Figura 1: Diagrama do fluxo de dados de cada usuário.

No **Experimento 2**, são utilizados **dois vetores de características distintos**, cada um representado em duas versões: **absoluta** e **proporcional**, com o objetivo de modelar duas bases de dados de naturezas diferentes: a *Base Simples* e a *Base Composta*.

- *Vetor da Base Simples (sem URL):* cada usuário é representado por um vetor de características contendo as aplicações utilizadas e o tempo total de uso em cada aplicação, de forma análoga à representação adotada no Experimento 1.
  Representação absoluta:
  $$\mathbf{v}^{abs}_{\text{simples}} = [1200,\ 300,\ 0]$$
  Representação proporcional:
  $$\mathbf{v}^{prop}_{\text{simples}} = [0.8,\ 0.2,\ 0]$$
- *Vetor da Base Composta (com URL):* como o conjunto de dados da Base Composta não possui informações de tempo de uso, o vetor é construído a partir do número de ocorrências da aplicação nos registros e da quantidade de domínios distintos acessados por meio de cada aplicação.
  Representação absoluta:
  $$\mathbf{v}^{abs}_{\text{composta}} = [(40,\ 10),\ (15,\ 4),\ (0,\ 0)]$$
  Representação proporcional:
  $$\mathbf{v}^{prop}_{\text{composta}} = \left[\frac{40}{55}, \frac{10}{14}, \frac{15}{55}, \frac{4}{14}, 0, 0\right]$$

4.4 Similaridade de Cosseno e Distância de Cosseno

A **Similaridade do Cosseno** mede o grau de alinhamento entre dois vetores, sendo útil para identificar quão semelhantes são os padrões de uso entre diferentes períodos.

Dessa forma, seja $\vec{u}$ o vetor referente ao vetor médio e $\vec{v}$ a um dos vetores diários do bloco de teste, a Similaridade do Cosseno entre dois vetores $\vec{u}$ e $\vec{v}$ é calculada pela fórmula:

$$\cos\_\text{sim}(u, v) = \frac{u \cdot v}{\|u\| \cdot \|v\|}$$

Para vetores com entradas não negativas o valor da Similaridade do Cosseno varia entre 0 e 1. Um valor próximo de 1 indica que os dois vetores apresentam padrões de uso semelhantes, enquanto valores próximos de 0 indicam padrões significativamente distintos. A **Distância de Cosseno** é sua forma complementar e expressa diretamente o grau de dissimilaridade entre vetores, definida como $d_{cos}(u, v) = 1 - \cos\_\text{sim}$. Valores próximos de 1 indicam maior divergência entre os perfis comparados.

4.5 Segmentação Temporal

Para avaliar a estabilidade temporal dos perfis, os dados de cada usuário foram divididos em dois períodos: treino e teste. A data de corte foi definida individualmente, considerando o número total de dias com registros de cada usuário. Os vetores baseados nas 50 aplicações mais usadas foram segmentados temporalmente em cinco proporções como mostra a Tabela 1:

Tabela 1: Proporções utilizadas na divisão treino–teste.

	10–90	**25–75**	**50–50**	**75–25**	**90–10**
Treino (%)	10	25	50	75	90
Teste (%)	90	75	50	25	10

Essas variações permitem analisar como a quantidade de dados de treino afeta a similaridade do cosseno entre o vetor médio do período de treino e o comportamento subsequente.

4.6 Filtros de Engajamento

Para garantir a qualidade da análise, foram aplicados filtros de engajamento: selecionaram-se apenas usuários com pelo menos 60 dias de atividade e que utilizaram no mínimo 5 das 50 aplicações mais frequentes. Esses critérios buscam assegurar perfis comportamentais consistentes e diversificados.

Após a filtragem, restaram 20.048 usuários (de 41.859 iniciais), representando 2,48 bilhões de registros (89% do total). Isso demonstra que a maior parte da atividade concentra-se em usuários com alto engajamento.

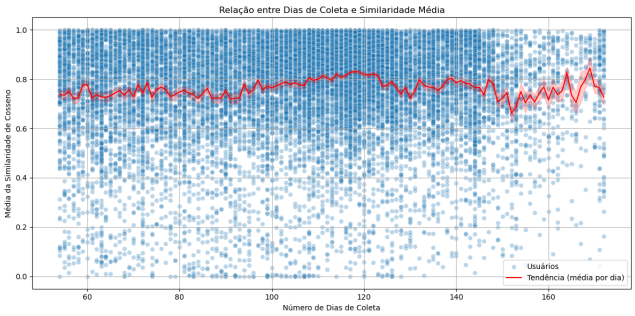

Figura 2: Relação entre dias de coleta e similaridade do cosseno na proporção 10/90 (treino/teste).

5 Experimentos

Nesta seção, são apresentados os resultados obtidos nos dois experimentos propostos, comparando o desempenho das medidas de similaridade em cenários com diferentes granularidades de dados.

5.1 Experimento 1: Generalização em Larga Escala

Este experimento foi projetado para responder à RQ1, avaliando se os perfis de uso de aplicações apresentam estabilidade temporal. Para cada usuário selecionado após os filtros de engajamento (Seção 4.5), foram construídos vetores diários $\mathbf{v}^d$ contendo o tempo total de uso nas 50 aplicações mais frequentes. Em seguida, a série temporal de cada usuário foi segmentada em períodos de treino e teste.O vetor médio de treino $\mathbf{m}$ foi calculado como a média dos vetores diários da janela inicial.

Cada ponto nos resultados representa a similaridade média individual de um usuário, calculada entre seus períodos de treino e teste. Por exemplo, um ponto em (x=60, y=0, 90) indica um usuário que possui 60 dias de atividade no bloco de treino e cuja média das similaridades entre o vetor médio de treino e os vetores diários de teste foi 0, 90. Esse valor indica alinhamento entre o padrão médio observado no período de treino e os padrões registrados posteriormente, evidenciando estabilidade temporal no comportamento vetorial desse usuário.

A Figura 2 mostra que, com apenas 10% da série temporal destinada ao treino, a similaridade média apresenta alta variabilidade entre os usuários. Isso ocorre porque o vetor médio é calculado a partir de uma janela muito curta e, portanto, é mais sensível a ruídos e dias atípicos. Mesmo assim, observa-se que, conforme o número de dias de coleta aumenta, a linha de tendência converge para valores próximos de 0, 78 e 0, 80, indicando que séries temporais mais longas levam a representações mais estáveis. Assim, mesmo sob baixa quantidade de treino, já se evidencia uma estrutura temporal consistente no comportamento dos usuários.

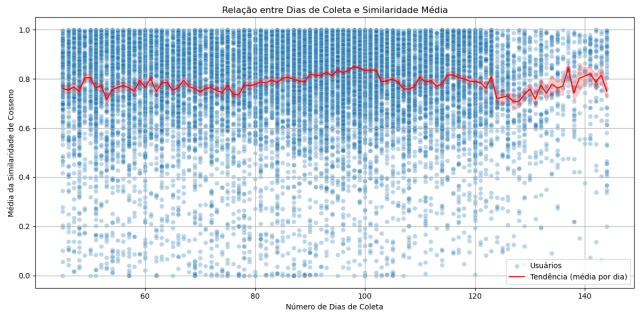

Figura 3: Relação entre dias de coleta e similaridade do cosseno na proporção 25/75 (treino/teste).

Na Figura 3, o aumento da janela de treino para 25% reduz significativamente a variabilidade das similaridades. O vetor médio passa a ser mais representativo e menos sujeito a variações, resultando em uma distribuição mais concentrada e em uma tendência mais estável, novamente próxima de 0, 80.A redução do limite superior no eixo X decorre da realocação de vetores diários do bloco de treino para o de teste. Assim, à medida que o bloco de teste aumenta, diminui a quantidade de dados disponível para o treino.

A Figura 4 mostra uma redução clara na variabilidade das similaridades em comparação às proporções menores. Com metade da série temporal sendo utilizada para compor o vetor médio de treino,

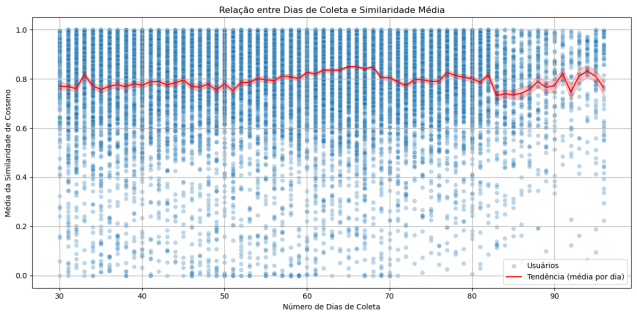

Figura 4: Relação entre dias de coleta e similaridade do cosseno na proporção 50/50 (treino/teste).

esse representante é significativamente mais robusto, o que reduz a influência de variações pontuais no comportamento do usuário. A linha de tendência permanece próxima de 0, 80 ao longo de todo o intervalo de dias de coleta, evidenciando que, mesmo entre usuários com séries relativamente curtas, a similaridade entre os períodos de treino e teste se mantém elevada e estável. Esse resultado indica que, a partir dessa proporção, o vetor médio é suficientemente expressivo para capturar o padrão comportamental individual com alta consistência.

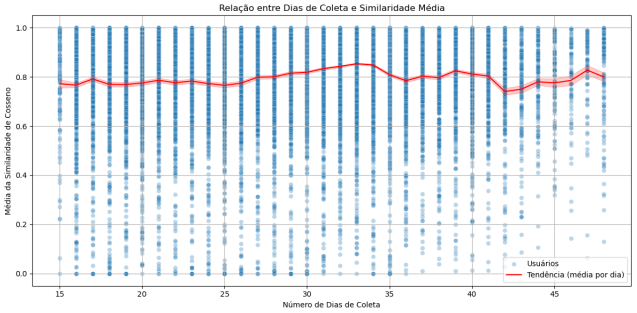

Figura 5: Relação entre dias de coleta e similaridade do cosseno na proporção 75/25 (treino/teste).

Já as Figuras 5 e 6 mostram respectivamente os resultados das proporções 75/25 e 90/10. Os resultados demonstram que o aumento da janela de treino conduz a representações progressivamente mais estáveis e precisas dos perfis de uso. Ambas confirmam que, quando o vetor médio é construído com uma parcela significativa do histórico, a estabilidade temporal torna-se marcante, reforçando a conclusão de que os perfis comportamentais dos usuários preservam uma estrutura consistente ao longo do tempo.

Nas proporções 75/25 e 90/10, a redução do número de dias disponíveis para o período de teste impõe algumas limitações interpretativas. Testes muito curtos tornam a média das similaridades menos sensível a mudanças reais no comportamento, além de reduzirem o número de usuários para a análise, como mostra a Tabela 2. Uma vez que séries temporais mais longas se tornam necessárias para atender aos requisitos de segmentação. Além disso, o vetor médio passa a incorporar quase todo o histórico, produzindo

valores de similaridade naturalmente mais estáveis. Embora isso não comprometa a avaliação da estabilidade temporal, mas parte dessa estabilidade decorre da dominância do período de treino.

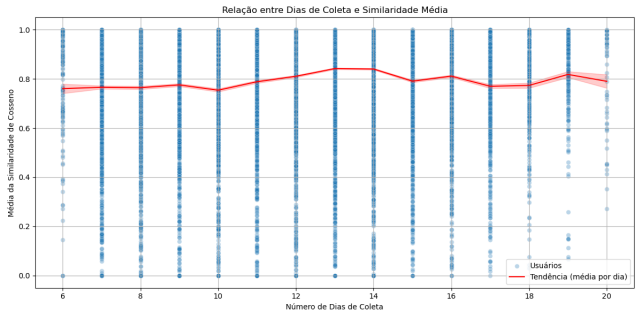

Figura 6: Relação entre dias de coleta e similaridade do cosseno na proporção 90/10 (treino/teste).

Tabela 2: Similaridade média e média de dias de teste em diferentes cortes

Corte	Média de Similaridade	Média de Dias de Teste por Usuário
10/90	0,7773	106,58
25/75	0,7951	88,82
50/50	0,8048	59,21
75/25	0,8089	29,86
90/10	0,8057	12,25

5.2 Experimento 2: Perfil Enriquecido com Navegação Web

Para aprofundar a comparação entre a Base Simples, que contém apenas por dados de uso de aplicações, e a Base Composta, que inclui informações de navegação, foi realizado um experimento integrando a Distância de Cosseno com duas formas complementares de avaliação: uma análise numérica de separabilidade e uma análise geométrica por meio de PCA. O objetivo foi verificar em que medida a presença de informações de URL altera a relação de proximidade entre o **usuário base** e as demais pessoas consideradas semelhantes na primeira etapa do estudo.

O **Experimento 2** busca avaliar a consistência da análise de similaridade quando se utiliza um novo tipo de dado comportamental, incorporando informações de navegação por meio de URLs, permitindo investigar a RQ2. O experimento foi estruturado em duas partes principais: (i) identificação dos vizinhos mais próximos na Base Simples e (ii) validação desses vizinhos em um cenário compatível com a Base Composta, construída com dados sintéticos.

5.2.1 Identificação dos Vizinhos na Base Simples. Na primeira etapa, foi utilizada exclusivamente a **Base Simples**, que não contém informações de URL. Nessa base, cada usuário é representado por um vetor de características composto pelas aplicações utilizadas e pelo tempo total de uso em cada aplicação.

A partir desses vetores, foram identificados os **10 usuários mais próximos do usuário base** por meio da métrica de Distância do Cosseno. Para essa etapa, foram avaliadas duas versões da representação: uma versão com dados absolutos e outra com dados proporcionais, conforme detalhado na Seção 4.2.

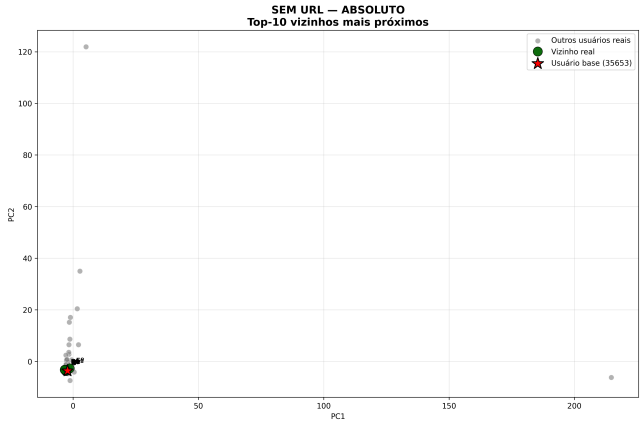

Figura 7: Análise de similaridade utilizando PCA para a Base Simples com representação absoluta.

Os resultados das projeções em PCA presentes nas Figuras 7 e 8 mostram que a representação proporcional tende a preservar melhor a organização estrutural dos perfis, produzindo maior dispersão entre os vizinhos e evidenciando diferenças sutis na forma da distribuição. Em contraste, a representação absoluta, embora realce diferenças de magnitude, resulta em aglomeração mais densa no espaço projetado, sugerindo menor poder discriminativo geométrico. Esse contraste explica os valores observados nas distâncias numéricas da Tabela 3 a média mais alta na representação absoluta não implica maior separabilidade, mas reflete a expansão artificial do espaço ocasionada pelo z-score.

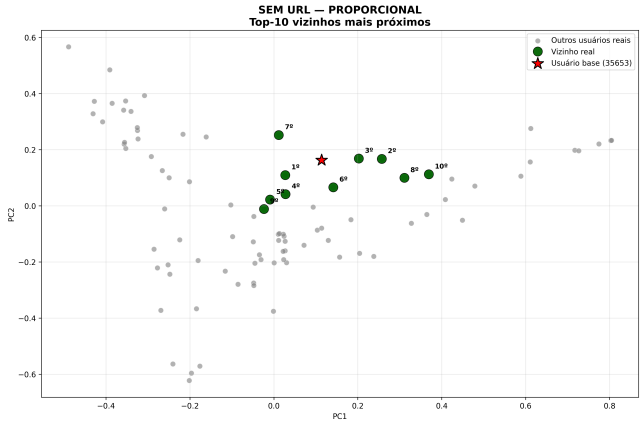

Figura 8: Análise de similaridade utilizando PCA para a Base Simples com representação proporcional.

5.2.2 Construção da Base Composta com Usuários Sintéticos. Para tornar possível a comparação entre as duas bases, foi necessária a reconstrução dos perfis correspondentes aos 10 vizinhos mais

próximos do **usuário base** identificados na primeira etapa. Como apenas o **usuário base** possui registros reais de navegação, esses perfis foram sintetizados de modo a preservar, de forma estatisticamente consistente, as principais características comportamentais dos usuários originais. O processo buscou manter a distribuição relativa de uso das aplicações, respeitar a estrutura individual de cada vizinho e incorporar padrões de navegação compatíveis com aqueles observados no **usuário base**. Com essa reconstrução, foi possível aplicar a mesma metodologia de comparação na **Base Composta** e analisar a relação de vizinhança sob a presença dos sinais adicionais de URLs.

Durante a geração dos perfis sintéticos, a foi adotado uma estratégia para evitar que esses usuários se tornassem artificialmente próximos ou excessivamente distantes do usuário base. Para isso, foram introduzidas variações controladas no número total de eventos por meio de fatores de escala aleatórios, e empregadas distribuições não uniformes na alocação desses eventos entre as aplicações, de modo a preservar a heterogeneidade típica observada em cenários reais. Assim como, garantir apenas uma sobreposição parcial entre o conjunto de aplicações dos perfis sintéticos e aquele utilizado pelo usuário base, evitando que a proximidade fosse inflada pela coincidência completa dos mesmos aplicativos. No componente de navegação, os vetores foram construídos a partir de uma combinação entre URLs reais observadas no **usuário base** e URLs sintéticas geradas de forma a manter qualidade estatística. Em conjunto, esses mecanismos permitiram produzir perfis diversificados, mas ainda coerentes com a estrutura comportamental esperada, assegurando que a análise posterior refletisse variações naturais do ambiente e não artefatos da construção dos dados.

5.2.3 Análise Final de Similaridade na Base Composta. Com os vetores reconstruídos para a Base Composta, foram novamente realizadas as projeções em PCA utilizando as representações absoluta e proporcional, permitindo avaliar como a inclusão dos sinais de navegação altera a estrutura de vizinhança originalmente observada na Base Simples. Os resultados dessas projeções são apresentados nas Figuras 9 e 10.

Na representação absoluta (Figura 9), é possível observar um espalhamento mais pronunciado dos vizinhos quando comparado à Base Simples. Esse aumento de dispersão indica que a incorporação dos eventos de navegação introduz novas dimensões comportamentais que ampliam o contraste entre os perfis. Embora a padronização por z-score contribua para a expansão do espaço vetorial, a separação adicional entre os usuários mostra a heterogeneidade introduzida pela diversidade de domínios e URLs associadas ao uso de aplicações.

A representação proporcional (Figura 10) mostra o impacto da navegação na distinção entre perfis. A normalização reduz o efeito da magnitude absoluta dos vetores, destacando relações angulares que capturam apenas a distribuição relativa entre aplicações e domínios. Mesmo sob esse processo de normalização, os vizinhos do usuário base passam a ocupar regiões mais distantes entre si e em relação ao próprio usuário base, mostrando que os padrões de navegação são fortemente discriminativos, independentemente do volume de uso.

Em conjunto, as projeções em PCA demonstram que a Base Composta possui maior poder de identificar que a Base Simples, tanto na forma absoluta quanto proporcional. A reorganização espacial observada nas Figuras 9 e 10 confirma que incluir informações de URLs altera a estrutura geométrica dos perfis, aumentando sua separabilidade e modificando a relação de vizinhança previamente identificada. Esses resultados reforçam que os sinais de navegação constituem uma dimensão comportamental adicional capaz de distinguir usuários que, considerando apenas o uso de aplicações, aparentavam ser muito semelhantes.

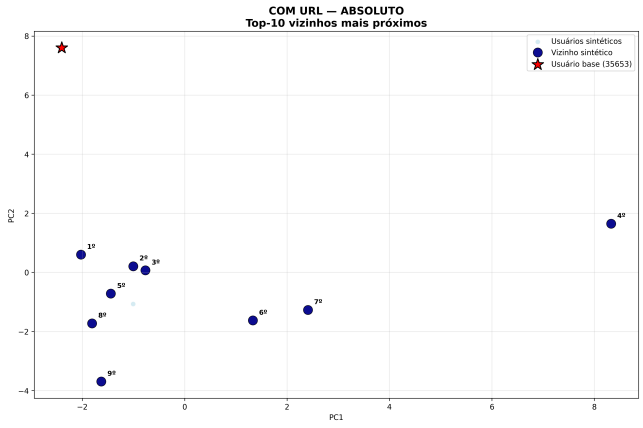

Figura 9: Análise de similaridade utilizando PCA para a Base Composta com uso de URL e representação absoluta.

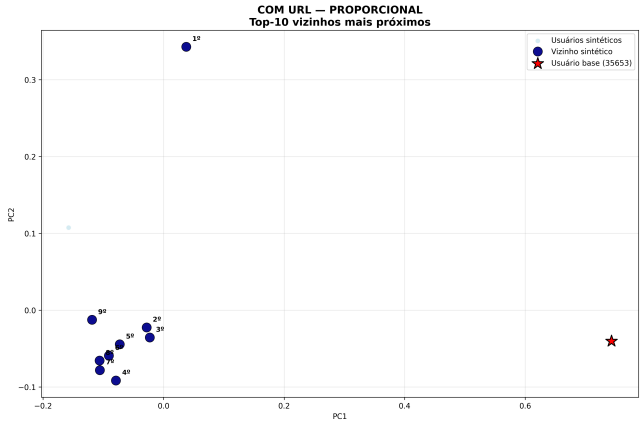

Figura 10: Análise de similaridade utilizando PCA para a Base Composta com uso de URL e representação proporcional.

5.2.4 Resultados Numéricos de Separabilidade. As Tabelas 3 e 4 apresentam as medidas de separabilidade dos 10 vizinhos mais próximos nos cenários avaliados, considerando as representações de vetores: absoluta e proporcional.

Os resultados numéricos complementam as projeções em PCA ao quantificar diretamente o impacto da inclusão das URLs na separabilidade entre o usuário base e seus vizinhos. Na Base Simples, as distâncias médias são relativamente baixas, refletindo a proximidade estrutural entre os perfis quando apenas o uso de aplicações é

considerado. Esse comportamento é esperado em ambientes corporativos, nos quais grande parte dos usuários compartilha rotinas e ferramentas semelhantes, resultando em vetores alinhados.

Ao reconstruir os mesmos perfis na Base Composta, observa-se aumento expressivo das distâncias em ambas as representações. Na forma absoluta, o crescimento reflete a diversidade adicional introduzida pelos eventos de navegação; na proporcional, o efeito é ainda mais pronunciado, pois a inclusão de URLs adiciona novos eixos de variação que alteram substancialmente as relações angulares entre os perfis. Como resultado, a estrutura de similaridade da Base Simples não se preserva.

Esses achados evidenciam o poder discriminativo dos padrões de navegação, promovendo maior separabilidade entre os usuários e modificando as relações de vizinhança, o que contribui diretamente para a resposta da RQ2.

Tabela 3: Separabilidade dos 10 vizinhos na representação absoluta (Z-score).

Base Absoluta	Média	Desvio	Mediana	Mínima
SEM_URL_ABS	0,8394	0,0417	0,8601	0,7225
COM_URL_ABS	1,1048	0,0565	1,1273	0,9704

Tabela 4: Separabilidade dos 10 vizinhos na representação proporcional (normalizada).

Base Proporcional	Média	Desvio	Mediana	Mínima
SEM_URL_PROP	0,0595	0,0267	0,0680	0,0168
COM_URL_PROP	0,5758	0,0502	0,5788	0,4914

6 Conclusão

Este trabalho apresentou o UserPrint, uma abordagem de perfilização comportamental baseada em vetorização de uso de aplicações e métricas angulares, avaliando sua capacidade de capturar estabilidade temporal e distinguir usuários em ambientes corporativos. Por meio de dois experimentos complementares, investigou-se a estabilidade individual (RQ1) e o impacto da inclusão de navegação web na separabilidade entre perfis (RQ2). No Experimento 1, foi possível observar que os perfis de uso de aplicações são estruturalmente estáveis ao longo do tempo. Mesmo com janelas curtas de treino, a similaridade média permanece elevada, e tende a convergir para aproximadamente 0, 8 em séries mais longas. Esse resultado indica que o comportamento funcional dos usuários possui regularidade suficiente para sustentar mecanismos de autenticação contínua ou detecção de desvios. Já o Experimento 2 mostrou que a incorporação de dados de navegação altera a estrutura de vizinhança entre perfis, aumentando a separabilidade tanto nas representações absolutas quanto proporcionais. As projeções em PCA e os valores numéricos de distância demonstraram que as URLs introduzem variabilidade adicional que não está presente nos vetores baseados apenas em aplicações, reforçando o caráter complementar entre estabilidade funcional e individualidade operacional. Os resultados sugerem que a combinação de múltiplas modalidades comportamentais amplia o poder discriminativo de modelos de análise de usuários, oferecendo uma base conceitual e empírica para sistemas de segurança adaptativos e contínuos. A necessidade de utilização de dados sintéticos na Base Composta para viabilizar a comparação com os vizinhos do usuário base, é uma limitação desse trabalho, porque restringe a generalização direta dos resultados desse experimento. Como trabalhos futuros, pretende-se ampliar a coleta com múltiplos usuários completos na base de navegação, a avaliação com janelas temporais deslizantes e a integração da abordagem UserPrint em sistemas reais de *User and Entity Behavior Analytics* (UEBA).

Agradecimentos

Este trabalho foi apoiado por: Bluepex, Laboratório de Pesquisa em Segurança, Engenharia de dados, Criptografia, Redes, Exploração e Testes (SECRET/UFPR), e Centro de Computação Científica e Software Livre (C3SL/UFPR), bem como pelo CNPq por meio de bolsa de produtividade em pesquisa.

Referências

[1] Jie Yang, Yuanyuan Qiao, Xinyu Zhang, Haiyang He, Fang Liu, and Gang Cheng. Characterizing user behavior in mobile internet. *IEEE Transactions on Emerging Topics in Computing*, 3(1):95–106, 2015. doi: 10.1109/TETC.2014.2381512.

[2] Bassam Sayed, Issa Traoré, Isaac Woungang, and Mohammad S. Obaidat. Biometric authentication using mouse gesture dynamics. *IEEE Systems Journal*, 7(2): 262–274, 2013. doi: 10.1109/JSYST.2012.2221932.

[3] Upal Mahbub, Jukka Komulainen, Denzil Ferreira, and Rama Chellappa. Continuous authentication of smartphones based on application usage. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 1(3):165–180, 2019. doi: 10.1109/TBIOM.2019.2918307.

[4] Heng Zhang, Vishal M. Patel, Mohammed Fathy, and Rama Chellappa. Touch gesture-based active user authentication using dictionaries. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 207–214, 2015. doi: 10.1109/WACV.2015.35.

[5] Alberto Urueña López, Fernando Mateo, Julio Navío-Marco, José María Martínez-Martínez, Juan Gómez-Sanchís, Joan Vila-Francés, and Antonio José Serrano-López. Analysis of computer user behavior, security incidents and fraud using self-organizing maps. *Computers & Security*, 83:38–51, 2019. doi: 10.1016/j.cose.2019.01.009.

[6] Rushit Dave, Marcho Handoko, Ali Rashid, and Cole Schoenbauer. From clicks to security: Investigating continuous authentication via mouse dynamics, 2024. URL https://arxiv.org/abs/2403.03828.

[7] Soumik Mondal and Patrick Bours. A study on continuous authentication using a combination of keystroke and mouse biometrics. *Neurocomputing*, 230:1–22, 2017. doi: 10.1016/j.neucom.2016.11.031.

[8] Salman Khaliq, Zain Ul Abideen Tariq, and Ammar Masood. Role of user and entity behavior analytics in detecting insider attacks. In *International Conference on Cyber Warfare and Security (ICCWS)*, pages 1–6, 2020. doi: 10.1109/3ICT.2018.8855807.

[9] Pierpaolo Artioli, Antonio Maci, and Alessio Magrì. A comprehensive investigation of clustering algorithms for user and entity behavior analytics. *Frontiers in Big Data*, 7, 2024. doi: 10.3389/fdata.2024.1375818.

[10] JinHua Xu and Hong Liu. Web user clustering analysis based on kmeans algorithm. In *International Conference on Information, Networking and Automation (ICINA)*, volume 2, pages V2–6–V2–9, 2010. doi: 10.1109/ICINA.2010.5636772.

[11] Charu C. Aggarwal. On the surprising behavior of distance metrics in high dimensional space. *Journal of Machine Learning Research*, 14:1–38, 2013.

[12] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations*, 2013.

[13] Haneen Arafat Abu Alfeilat, Ahmad BA Hassanat, Omar Lasassmeh, Ahmad S Tarawneh, Mahmoud Bashir Alhasanat, Hamzeh S Eyal Salman, and VB Surya Prasath. Effects of distance measure choice on k-nearest neighbor classifier performance: a review. *Big data*, 7(4):221–248, 2019.