

Hagrid: A Modular AI Pipeline for Cookieless User Profiling in Digital Publishing

Luca Aliberti¹, Francesco Apicella², Martina Buttarelli², Luca Carrubbo¹, Camilla Murati^{2,*} and Davide Perfetto²

¹University of Salerno, Via Giovanni Paolo II 132, 84084 Fisciano (SA), Italy

²Evolution Group, Italy

Abstract

The progressive removal of third-party cookies as a cross-site tracking mechanism and the tightening of the European regulatory framework (GDPR, DSA, AI Act) are driving digital publishing towards profiling solutions that are effective, privacy-preserving, and transparent. This paper presents Hagrid, a modular AI platform for context-aware user profiling designed for a network of more than 400 Italian online publishers. Hagrid is structured as a three-stage pipeline comprising data collection, content classification, and user profiling. The profiling layer combines representation learning with Factorization Machines and unsupervised clustering to derive interpretable behavioural archetypes from interaction data and semantic content labels. It operates across three contextual dimensions (User Agent, Time, and Content) and leverages Context Space Theory (CST) to organize the latent space into semantically meaningful regions. Experiments on real-world navigation data show that Factorization Machines match deeper architectures such as xDeepFM in predictive quality while reducing training time by up to 16×, with peak incremental performance reaching HR@10 = 0.928 and NDCG@10 = 0.526. The resulting archetypes capture interpretable consumption patterns and enable effective warm-start profiling under cold-start conditions typical of cookieless traffic.

Keywords

User Profiling, Representation Learning, Factorization Machines, Context Space Theory, Cookieless Web, Digital Publishing

1. Introduction

The technical foundations of user profiling in digital publishing are being reshaped by two concurrent forces: the progressive removal of third-party cookies as a cross-site tracking mechanism and the consolidation of the European regulatory framework through GDPR, the Digital Services Act (DSA), and the AI Act [1, 2, 3]. As third-party cookie blocking becomes the default for a growing share of users, advertising platforms are shifting their tracking strategies [4]. At the same time, regulatory requirements on data minimization, transparency and algorithmic accountability apply to every system that processes behavioural signals at scale.

Traditional context-aware profiling systems incorporate contextual factors such as time, location or device into the personalization process [5, 6], but typically assume the availability of a stable user identity across sessions. When these identifiers are removed, the profiling system must rely on contextual signals such as the type of device, the time of access, and the semantic category of the content being consumed. Building meaningful user profiles from these pseudonymous, session-level signals poses specific challenges that go beyond what standard profiling approaches are designed to handle. Cold start is not an edge case but the default operating condition, since every browsing session can appear as a new user. In the Evolution Adv network¹, which serves the publisher base on which Hagrid operates, approximately 75% of traffic originates from mobile devices, browsing sessions are short, and user turnover is high: this means that the profiling system must be able to produce a meaningful representation from the very first interaction. At the same time, the profiles must be interpretable

WOA 2026: 27th Workshop "From Objects to Agents", 2026, Italy

*Corresponding author

✉ camilla.murati@evolutiongroup.it (C. Murati)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

¹<https://evolutiongroup.digital>

enough for editorial teams to understand what the resulting audience segments actually represent. Privacy compliance cannot be a post-hoc layer but must be structural, built into the way data flows through the system from the first point of collection.

This paper presents Hagrid, a modular AI platform for user profiling in digital publishing without persistent identifiers. The platform is organized as a pipeline of three processing stages, each responsible for a distinct function: data acquisition with fingerprinting-based identification [7, 8] and non-reversible hashing across a network of more than 400 Italian publisher sites, semantic content classification according to the IAB Content Taxonomy [9], and user profiling through representation learning and unsupervised clustering. The core AI contribution of this paper lies in the profiling stage, which combines a Factorization Machine (FM) [10], a representation learning model that projects users, content and context into a shared latent space, with Context Space Theory (CST) [11, 12], which provides a formal semantic structure over the learned representations so that the output of the learning process is not only accurate but also interpretable by domain experts.

The contribution of this paper is twofold. First, it describes the architecture of the Hagrid platform and the design choices that are driven by the absence of persistent identifiers and by the regulatory requirements that apply to a live publisher network. Second, it provides a detailed experimental evaluation of the profiling component on real navigation data collected from the publisher network. The results show that the FM matches xDeepFM in predictive quality at approximately one sixteenth of the training cost, that the archetype-based warm-start mechanism addresses the cold-start problem from the very first interaction, and that the model does not degrade catastrophically under distributional drift, with performance peaking at a moderate data volume.

The remainder of the paper is organized as follows. Section 2 reviews the relevant literature. Section 3 presents the Hagrid platform architecture. Section 4 reports the experimental evaluation. Section 5 discusses the results and their implications. Section 6 concludes the paper and outlines directions for future work.

2. Related Work

User profiling has progressively moved from static, keyword-based representations to dynamic models that learn dense embeddings from interaction data [13, 14]. Alternative approaches model preferences through interpretable structures such as Fuzzy User Signatures [15], which represent user activities through fuzzy relations and compute similarities via asymmetric kindredness measures. This evolution, however, has developed almost entirely under the assumption that a stable user identifier is available across sessions, enabling the system to accumulate a behavioural history over time. The deprecation of third-party cookies disrupts this assumption for a growing share of web traffic. Browser fingerprinting [7, 8] provides a stateless identification mechanism by combining technical signals from the browser environment into a probabilistic device signature, but its integration into a full profiling pipeline has received limited attention. Other strategies operate at a coarser level: the Topics API [16] assigns broad interest labels on-device, federated approaches [17] keep data on the client, and contextual targeting bypasses user-level signals entirely. None of these supports the construction of fine-grained, server-side user profiles from real navigation data. The gap extends to evaluation resources: public benchmarks such as MIND [18] and Adressa [19] rely on cookie-based or login-based identification on single editorial platforms, leaving the multi-publisher, fingerprinting-based setting unaddressed.

In such a setting the profiling system must rely on contextual signals such as the technical characteristics of the access (device, operating system, browser), the time of the interaction, and the semantic category of the content being consumed. A model capable of learning from sparse, high-dimensional categorical inputs without manual feature engineering is therefore essential. Factorization Machines [10] meet this requirement by factorizing all pairwise feature interactions as inner products of low-dimensional latent vectors [20]. The literature has predominantly employed them, alongside deep extensions such as DeepFM [21], xDeepFM [22] and Neural Factorization Machines [23], as prediction engines for item-level recommendation: the model receives a user–item pair and outputs a

relevance score. Comparative evaluations in the news domain [24, 25] have shown that the deeper variants offer marginal accuracy gains over the standard FM at substantially higher computational cost. Hagrid adopts the FM and uses the resulting latent space not as the endpoint but as the substrate for profile construction: the embeddings are decomposed into contextual dimensions and aggregated into a normalized profile vector that captures who the user is at a given moment.

This shift in usage, however, does not by itself guarantee that the resulting profiles are interpretable. Without additional structure, the latent space learned by the FM is a numerical artefact in which proximity reflects statistical co-occurrence but carries no semantic meaning that domain experts can inspect. CST [11] addresses this limitation by defining a context space as a product of typed dimensions and interpreting situations as regions in that space. Extensions of CST to runtime context prediction [26], to machine learning settings [12], and to situation-aware monitoring of complex domains [27] have demonstrated that imposing this formal structure on learned representations enables semantic reasoning over the latent space. A related direction has explored the integration of Situation Awareness theory with FM-based architectures in news recommendation [25], showing that aggregating contextual features into explicit situational representations improves ranking quality and interpretability. Hagrid adopts a complementary perspective in the profiling domain: CST decomposes the FM latent space into named contextual dimensions, defines profiles as directions rather than points, and identifies behavioural archetypes as stable, semantically readable regions of the space.

3. The Hagrid Platform

Hagrid is organized as a modular pipeline designed to address three operational constraints that arise from operating a profiling system across a live network of more than 400 publisher sites. First, the absence of persistent user identifiers, driven by the deprecation of third-party cookies and reinforced by the privacy-by-design choice of replacing any potentially identifying field with a non-reversible hash at ingestion time, requires that all representations be derived from pseudonymous, context-level signals. Second, the system is designed so that each processing stage can be inspected independently, which is a practical requirement when operating on behavioural data at scale [1, 3]. Third, the editorial mix of the network shifts continuously, requiring that individual components be updated on independent schedules without disrupting the rest of the pipeline.

The platform comprises three processing stages, illustrated in Figure 1: (i) an acquisition layer that ingests navigation events, (ii) a classification layer that assigns semantic labels to editorial content, and (iii) a profiling layer that derives user profiles and behavioural archetypes. The acquisition layer feeds both downstream stages, while the classification layer supplies the semantic labels that constitute the Content dimension of the user profiles.

3.1. Acquisition Layer

The acquisition layer is the platform’s interface with the external environment. It is built on a three-tier architecture that separates data collection from asynchronous processing and downstream consultation.

The first tier observes raw navigation events generated by a JavaScript collection endpoint integrated into the advertising script that runs on every publisher page in the Evolution Adv network. Each event carries the visited URL, a precise timestamp, and a set of technical metadata derived from the browser environment (device type, operating system, browser family, device brand, screen resolution). User identification follows a stateless approach based on browser fingerprinting [7, 8]: instead of relying on cookies, the system combines multiple technical signals from the browser environment to construct a composite device signature. This signature is immediately transformed through non-reversible cryptographic hash functions (SHA-256) into two hashed tokens that replace the original signals before any data is persisted or transmitted to downstream components. The first token is a session-level identifier computed from the concatenation of the user agent string and a partial IP address with a daily salt, providing stable tracking within a single day and site. The second token is a probabilistic cross-session identifier derived from a broader set of fingerprinting signals, designed to

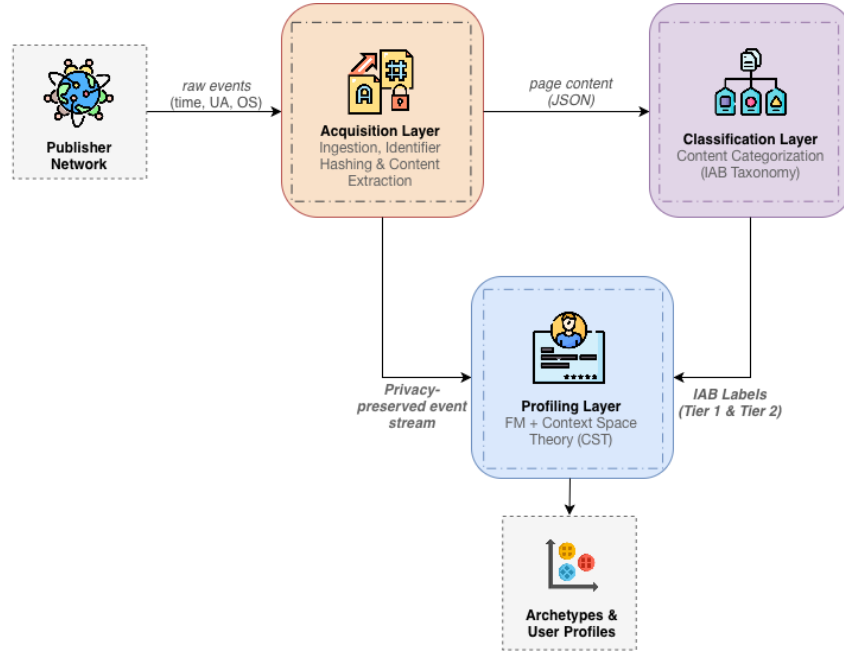


Figure 1: Overview of the Hagrid platform architecture. The acquisition layer ingests navigation events and extracts page content. The classification layer categorizes content against the IAB Taxonomy. The profiling layer receives both the event stream and the semantic labels to derive behavioural archetypes through FM and CST.

maintain consistency across sites and over a longer temporal horizon (up to fourteen days), although its stability is not guaranteed. Neither token is reversible to the original signals, and no identifier in its original form is ever written to disk. This non-reversible transformation at the point of ingestion is consistent with the broader privacy-preserving design principles for user activity analysis discussed in [28], where cryptographic mechanisms are used to anonymize interactions while preserving the analytical utility of the resulting signals.

It should be noted that, under the GDPR, hashed fingerprint-derived tokens are more appropriately regarded as pseudonymous rather than anonymous identifiers. Even when cryptographic hashing is applied, a stable identifier may remain susceptible to re-identification through correlation with auxiliary information or repeated observations. Three safeguards are therefore applied at ingestion time: the short-term token is re-salted daily and cannot be linked across days; the long-term token is retained for at most fourteen days; and the original fingerprinting signals are discarded immediately after token generation, making reconstruction of the original feature set computationally infeasible. These measures substantially reduce re-identification risks and align the platform with the GDPR notion of pseudonymization, without claiming full anonymization.

The second tier handles the computationally intensive operations that cannot be performed synchronously at ingestion time. The page URL is dispatched to a system of specialized asynchronous queues, each dedicated to a distinct processing category: incoming event registration, content extraction through web scraping, and data aggregation. This separation ensures that variability in the response times of external publisher servers does not affect the latency or reliability of the collection endpoint. The content-extraction subsystem retrieves the textual body of each visited page and stores it in structured JSON format together with its metadata.

The third tier exposes the collected data through REST APIs, structured exports, and a web interface (HagridUI) that provides search, filtering and visualization capabilities for the editorial team.

The acquisition layer produces two data products that feed the downstream stages: (i) an event stream recording the sequence of user interactions with timestamps and contextual metadata, consumed by the profiling layer; and (ii) a corpus of structured page contents in JSON format, consumed by the classification layer.

3.2. Classification Layer

The classification layer assigns a semantic label to every piece of editorial content that enters the platform. Each article collected by the acquisition layer is mapped to the IAB Content Taxonomy at both Tier 1 (30 main categories) and Tier 2 level [9]. The layer is designed as a modular and extensible component: the labelling logic is decoupled from the embedding model used to compute textual representations, so that the underlying model can be swapped without affecting the rest of the pipeline. The labelling process is based on semantic similarity in a shared embedding space. The title and body of each article are concatenated and encoded into a dense vector representation using a pre-trained Sentence Transformer [29]; in this work, the BAAI/bge-base-en-v1.5 model is adopted, which is optimized for semantic similarity and retrieval tasks, but any compatible encoder can be plugged in as the embedding backend. In parallel, the textual descriptions of all IAB taxonomy nodes (obtained by concatenating the Tier 1 and Tier 2 labels) are encoded with the same model. All vectors are L2-normalized so that cosine similarity coincides with the inner product. The normalized article embeddings are then compared against the full set of taxonomy embeddings using Facebook AI Similarity Search (FAISS) [30], which performs efficient nearest-neighbour search over the indexed taxonomy vectors. Each article is assigned the category and subcategory corresponding to the taxonomy node with the highest cosine similarity. Special rules handle edge cases such as homepage URLs and error pages, which are assigned dedicated labels to prevent misclassification. The output is a category and subcategory label that is persisted alongside the article metadata.

The classification task is non-trivial in this setting because the publisher network spans a wide editorial spectrum, from national news outlets to niche sports or lifestyle magazines. A single day of collection can include articles ranging from match reports to parliamentary coverage to recipe guides, and the classifier must assign each of them to the correct position in a taxonomy of over 400 subcategories. For example, an article covering a tennis match at Wimbledon is mapped to *Sports > Tennis*, so that all user interactions with that article carry a precise thematic signal into the profiling layer. The reliability of this mapping is critical: the Content dimension of the Context Space is built entirely from the labels produced by this module, and any systematic misclassification would propagate into the user profiles and ultimately into the behavioural archetypes. Without a reliable mapping from editorial content to a shared taxonomy, the profiling module would have no thematic signal to work with and the resulting archetypes would be limited to device and temporal patterns alone.

3.3. Profiling Layer

The profiling layer constructs user profiles and derives behavioural archetypes from the combination of interaction data and semantic content labels. From a machine learning perspective, the layer addresses two tasks in sequence: a representation learning task, in which a Factorization Machine [10] learns a shared latent space from sparse, high-dimensional categorical interaction data; and an unsupervised learning task, in which the learned representations are clustered to discover recurring behavioural patterns. The use of FM-based architectures for context-aware modelling on real navigation data has been investigated in related work on news recommendation [25], which provides supporting evidence on the suitability of FMs for this category of signals. Context Space Theory (CST) [11, 12] provides the bridge between these two tasks: it imposes a formal semantic structure on the latent space so that the output of the representation learning stage is not an opaque numerical artefact but a space in which distances, directions and regions carry interpretable meaning. In what follows, the term *latent space* refers to the numerical embedding produced by the FM, while *Context Space* denotes the same space once the CST decomposition into named dimensions and the normalization of Eq. (3) have been applied. The layer operates through a dual-loop architecture that separates the computationally heavy learning operations (offline) from lightweight real-time inference (online).

The layer receives two input streams: (i) the event stream from the acquisition layer, carrying user tokens, timestamps and technical metadata (device type, operating system, browser family, device brand); and (ii) the IAB labels from the classification layer, associating each visited page with a position

in the content taxonomy. Each interaction is therefore characterized by three contextual dimensions: User Agent, Time, and Content.

In the offline phase, an FM is trained on historical interaction data to learn a shared latent space. The FM models each interaction as a sparse feature vector encoding user, content and context variables, and computes a predictive score that combines a linear term with a bilinear interaction term. The interaction coefficients are factorized as inner products of low-dimensional latent vectors $\mathbf{v}_i \in \mathbb{R}^k$ associated with each feature i , which allows the model to generalize to rare or unseen feature combinations while keeping inference linear in the number of active features [10]. The model is trained using a session-level variant of Bayesian Personalized Ranking (BPR) [31], a learning-to-rank objective that constructs preference pairs from clicked and non-clicked items within each impression and applies softmax-derived weights so that harder cases receive greater emphasis during training.

The latent vectors learned by the FM are not treated as opaque numerical representations. CST provides both the interpretive framework and the formal structure for organizing them into a semantically coherent space. A Context Space is defined [11] as $CS = (D, R)$, where $D = \{d_1, \dots, d_n\}$ is the set of contextual dimensions and $R = \prod_i R_i$ is the product of their respective domains. In the Hagrid setting, the three principal dimensions are

$$D = \{UA, TIME, CNT\}, \quad (1)$$

where UA covers the technical access characteristics, $TIME$ covers the temporal patterns, and CNT covers the semantic content categories. Each interaction t is a point in this space, and its projection into the latent representation learned by the FM is computed as

$$\mathbf{z}_t = \alpha \mathbf{g}_{UA}(t) + \beta \mathbf{g}_{TIME}(t) + \gamma \mathbf{g}_{CNT}(t), \quad (2)$$

where each $\mathbf{g}(t)$ is the sum of the embedding vectors of the active features in the corresponding dimension group, and $\alpha, \beta, \gamma \geq 0$ are coefficients that control the relative emphasis of the three dimensions. This decomposition makes the contribution of each contextual dimension explicit and tunable, and it is the point where the FM representation acquires the semantic structure prescribed by CST: the resulting vector $\mathbf{z}_t$ is not an opaque numerical point but a position in a space whose axes have names and whose directions carry interpretable meaning.

The user profile is then obtained by aggregating the interaction signatures over a temporal horizon. For a user u with recorded interactions $I(u)$, the profile is

$$\mathbf{p}_u = \frac{\sum_{t \in I(u)} w_t \mathbf{z}_t}{\left\| \sum_{t \in I(u)} w_t \mathbf{z}_t \right\|}, \quad (3)$$

where $w_t \geq 0$ are temporal weights that can be uniform or recency-biased. The L2 normalization projects the profile onto the unit hypersphere, so that only the direction of the vector is retained, not its magnitude. This means that two users who share similar preferences but differ in activity volume will be geometrically close, while users with divergent interests will be distant regardless of how active they are. The inner product between two normalized profiles coincides with cosine similarity, providing a natural and interpretable distance metric in the Context Space.

Once the profiles are computed, the unsupervised learning stage begins. The set $\{\mathbf{p}_u\}$ is clustered using k -means with cosine distance to identify behavioural archetypes. Each cluster centroid

$$\mathbf{m}_c = \frac{1}{|\mathcal{U}_c|} \sum_{u \in \mathcal{U}_c} \mathbf{p}_u \quad (4)$$

represents the average direction of a group of users with similar consumption patterns and admits a direct editorial interpretation. In terms of CST, these centroids correspond to stable regions of the

context space, i.e., recurring configurations of device, temporal and thematic variables that characterize a recognizable type of content consumption.

In the online phase, the system generates or updates a profile for each incoming user. For known users, the profile is updated incrementally as new interactions arrive via Eq. (3). For new users, which represent the typical case in traffic without persistent identifiers, a provisional profile is assigned based on the nearest archetype centroid:

$$c(u_{\text{new}}) = \arg \min_k \| \mathbf{z}_{u_{\text{new}}} - \mathbf{m}_k \|, \quad (5)$$

so that a single interaction is sufficient to place the user in a meaningful region of the Context Space. As further interactions accumulate, the profile is recomputed and the archetype assignment is updated accordingly, allowing the representation to converge progressively towards the user’s actual behavioural pattern. In the downstream application, the profile is used to compute relevance scores over candidate items via the FM scoring function and produce a Top-K content recommendation.

The separation between offline and online loops reflects an architectural choice driven by operational constraints. The offline loop involves model retraining and re-clustering, which are computationally expensive but infrequent. The online loop relies exclusively on lightweight vector operations (embedding lookup, weighted averaging, nearest-centroid assignment), designed to meet the latency requirements of a real-time publishing environment. The choice of the FM over deeper alternatives such as xDeepFM [22] is motivated by this same constraint, as the evaluation in Section 4 confirms.

4. Experimental Evaluation

The experimental evaluation focuses on the profiling component of Hagrid. The acquisition and classification layers are validated only indirectly, as upstream sources of the data consumed by the profiling layer. Since profiling and downstream content recommendation share the same learned representations, the quality of Top-K recommendations is used as a proxy to measure how well the underlying embeddings capture behavioural patterns.

The evaluation investigates whether the FM is a competitive alternative to deeper models, how it performs under cold-start conditions that dominate traffic in the absence of persistent identifiers, and whether the resulting behavioural archetypes are semantically coherent and operationally useful.

4.1. Dataset and Protocol

All experiments are run on data acquired by the acquisition layer from four publisher domains within the Evolution Adv network, selected to span a broad editorial spectrum: a national news agency, a national newspaper, a regional newspaper, and a generalist online portal. The dataset spans a continuous observation window of twenty days (9–29 July 2025) and contains approximately 64,493 click events generated by 2,746 unique users on 4,688 distinct articles, partitioned into a training set (51,149 events, 2,555 users), a development set (7,482 events, 475 users), and a held-out test set (5,862 events, 408 users). Each event is enriched with contextual metadata and semantic labels produced by the classification layer.

Each interaction is characterized by attributes spanning three contextual dimensions: semantic content, technical access characteristics, and temporal dynamics. Table 1 summarises the extracted features and their domains.

The raw event stream is preprocessed as follows:

- Homepage accesses and URLs with a depth below 2 are excluded to filter out incidental traffic and focus on deliberate content choices. Duplicate events and consecutive involuntary refreshes are also removed.
- User-day pairs with fewer than 5 interactions are discarded to prevent statistical noise and spurious behaviors from skewing the representation.

Table 1

Overview of the contextual features extracted for each user interaction.

Dimension	Feature	Domain / Values
Content	IAB Taxonomy	Category and subcategory
	Publication	Timestamp
	URL Structure	Depth and path
User Agent (UA)	device_type	{mobile, tablet, desktop, unknown}
	os_family	{android, ios, windows, macos, linux, unknown}
	browser_family	{chrome, safari, firefox, edge, opera, samsung, unknown}
	brand	{samsung, huawei, xiaomi, apple, oppo, vivo, other, unknown}
Temporal	Hour of the day	Integer $\in [0, 23]$
	Time band	Categorical (6 bins)
	Day of the week	Categorical (7 days)
	Weekend indicator	Binary {0, 1}

- Events are ordered by user and grouped into sessions based on a 5-minute inactivity window. A new session begins whenever the gap between consecutive events exceeds 300 seconds or a day change occurs.

For the training phase, negative examples (articles available but not clicked) are sampled under strict, realistic constraints. A valid negative sample must satisfy three conditions:

- It must belong to the same publisher site and the same day as the positive event.
- It must have been published within the 3 hours preceding the click.
- It must not have been already viewed by the user.

Two complementary validation protocols are adopted to evaluate the model under realistic conditions. The first consists of a classic split, where sessions are sorted chronologically and partitioned into 80% training, 10% development, and 10% test sets, thereby preventing temporal leakage between training and evaluation.

The second protocol involves an incremental split designed to simulate the operational life cycle of the system. The initial training set is formed by the first 40% of the sessions, while the remainder is divided into ten contiguous temporal blocks. At each of five iterations, one additional block is accumulated in the training set, with the subsequent block serving as the development set and the following one as the test set. This protocol exposes the model to realistic distributional drift and tracks how training cost scales with data volume.

The reported metrics are Hit Ratio (HR@K), Mean Average Precision (MAP@K) [32], and Normalized Discounted Cumulative Gain (NDCG@K) [33], for $K \in \{1, 3, 5, 10\}$. These are standard top-K ranking metrics [34], used here to assess the quality of the latent representation that underlies the profiling layer rather than as evaluation criteria of a recommendation system. NDCG@3 is the principal metric for model selection, as it heavily penalizes representations that fail to place highly relevant content at the very top of the ranked list. Training time per epoch is also reported as a non-negligible operational cost.

4.2. Model Comparison

The first experiment compares the FM adopted in Hagrid against xDeepFM [22], a deep variant that adds a Compressed Interaction Network to the standard FM. Both models are trained with embedding dimension 64, learning rate 2×10^{-3} , L2 regularization 2×10^{-5} , and batch size 256, with early stopping on validation NDCG@3.

The results, summarized in Table 2, show that the two models are essentially equivalent in predictive quality across all metrics, with absolute differences of the order of a few hundredths of a point and no consistent direction of advantage. At NDCG@3, the FM scores 0.341 against 0.349 for xDeepFM. On HR@5 and HR@10, the FM is slightly ahead. These results are consistent with findings reported by [25, 24] on the MIND benchmark, where deeper architectures provide only marginal improvements over FM for news recommendation tasks.

The picture changes when considering training cost: the FM requires 52.65 ± 4.06 seconds per epoch against 835.08 ± 67.18 seconds for xDeepFM, at a comparable parameter count ($\sim 40.5\text{M}$ for both). The FM is approximately sixteen times faster to train at matching predictive quality, which justifies its adoption as the representational engine of the profiling layer: the offline retraining loop must be cheap enough to run as distributional drift accumulates, and the online inference loop must be light enough to provide low-latency profile updates.

Table 2

Predictive comparison between the FM adopted in Hagrid and xDeepFM. Bold marks the best value per row.

Metric	FM	xDeepFM	Δ
HR@1	0.1976	0.2202	+0.0226
HR@3	0.4501	0.4468	-0.0033
HR@5	0.6285	0.6042	-0.0243
HR@10	0.8649	0.8541	-0.0108
MAP@3	0.3035	0.3157	+0.0122
MAP@10	0.3761	0.3848	+0.0087
NDCG@3	0.3408	0.3492	+0.0084
NDCG@10	0.4911	0.4947	+0.0036
Time/epoch (s)	52.65 $\pm$ 4.06	835.08 $\pm$ 67.18	$\times 15.9$

4.3. Cold-Start Analysis

The second experiment addresses the cold-start condition. Users in the test set are split into *seen* users (present in the training set) and *unseen* users (entirely new). The results, summarized in Table 3, show a systematic degradation: HR@1 drops from 0.198 to 0.124, HR@5 from 0.629 to 0.458, and NDCG@3 from 0.341 to 0.230, with absolute losses ranging between -0.07 and -0.17 .

Table 3

Performance on seen versus unseen users.

Metric	Seen	Unseen	Δ
HR@1	0.1976	0.1241	-0.0735
HR@3	0.4501	0.3084	-0.1417
HR@5	0.6285	0.4580	-0.1705
HR@10	0.8649	0.7518	-0.1131
MAP@3	0.3035	0.2032	-0.1003
MAP@5	0.3446	0.2374	-0.1072
MAP@10	0.3761	0.2772	-0.0989
NDCG@3	0.3408	0.2301	-0.1107
NDCG@5	0.4146	0.2916	-0.1230
NDCG@10	0.4911	0.3873	-0.1038

This degradation is a structural limit of any model that depends on accumulated user history. It is this observation that motivates the warm-start mechanism described in Section 3.3: the nearest archetype centroid in the Context Space (Eq. (5)) provides a behaviourally plausible provisional profile from the very first interaction, and the feasibility of this assignment depends on the centroids being stable regions of a semantically structured space rather than arbitrary numerical points.

4.4. Incremental Evaluation

The third experiment runs the profiling layer under an incremental split designed to simulate the operational life cycle of the system. The initial training set is formed by the first 40% of the sessions, while the remainder is divided into ten contiguous temporal blocks. At each of five iterations, one additional block is accumulated in the training set, with the subsequent block serving as the development set and the following one as the test set. This protocol exposes the model to realistic distributional drift and tracks how training cost scales with data volume.

Table 4 reports the results across the five temporal versions. HR@10, MAP@10 and NDCG@10 improve from *v1* to *v3*, peaking at 0.928, 0.402 and 0.526 respectively, and then decline on *v4* and *v5*. Training time grows linearly with the size of the training set, from 27.3 s/epoch (*v1*) to 53.0 s/epoch (*v5*), confirming that the computational cost remains proportional and manageable.

Table 4

Incremental evaluation across five temporal versions.

Metric	v1	v2	v3	v4	v5
HR@10	0.477	0.568	0.928	0.809	0.842
MAP@10	0.164	0.200	0.402	0.287	0.325
NDCG@10	0.235	0.285	0.526	0.408	0.445
Time/epoch (s)	27.3	36.1	42.2	46.6	53.0

The peak at *v3* followed by a decline on *v4* and *v5* suggests that beyond a certain volume, the accumulation of temporally distant sessions introduces distributional noise that offsets the benefit of additional training data. The key observation is that the model reaches strong performance with a moderate amount of data (*v3*), confirming that the FM can be effectively retrained on rolling temporal windows rather than on the full historical archive. The linear growth in training time, combined with the speedup over xDeepFM reported in Section 4.2, is what makes periodic retraining sustainable under realistic operational cadences.

4.5. Behavioural Archetypes

The fourth analysis examines the behavioural archetypes emerging from clustering in the Context Space. Figure 2 reports the two-dimensional PCA projection of user profiles, with cluster assignments and centroids. Three archetypes are identified for subsequent analysis. The number of clusters was selected by evaluating internal validation metrics, including silhouette score and within-cluster sum of squares, over $k \in \{2, \dots, 8\}$. The resulting curves did not exhibit a clear optimum, suggesting a relatively smooth structure in the latent space, where user behaviour is better represented along continuous dimensions rather than sharply separated groups. In this setting, and in the absence of a strong quantitative preference for a specific k , the final choice was guided by interpretability considerations and validated in discussion with domain experts, selecting the smallest segmentation that preserved meaningful separation in the latent space while remaining actionable for editorial use.

The first archetype (29% of users) is characterized by a predominant interest in Sports (25%), with a notable concentration in Tennis (17%), alongside Sensitive Topics (24%). Access occurs primarily in the afternoon (14:00–17:00) with a balanced mix of mobile and desktop devices. In editorial terms, this group corresponds to an informed reader who follows sports events during the working day.

The second archetype (32%) shows a strong mobile orientation (80% of accesses) and a weekend concentration (55%). Thematic interests are balanced between Sensitive Topics (24%) and Sports (23%), with access peaks in the morning and early evening. This profile corresponds to a casual reader who consumes content during leisure time, predominantly on a smartphone.

The third archetype (39%) is desktop-dominant (60% of accesses) and almost entirely weekday-based (91%). Content preferences include Sensitive Topics (31%) and Sports (25%), with a distinctive presence

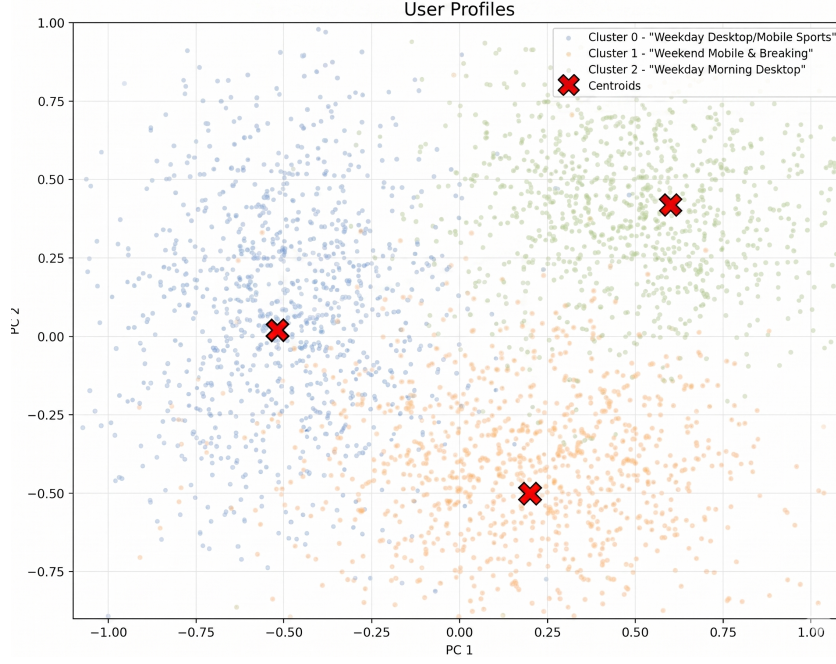


Figure 2: PCA projection of user profiles in the Context Space. Each point represents a user, colours indicate cluster membership, and crosses mark the archetype centroids.

of Fine Art and Opera (4.6%). Access is concentrated in the morning (8:00–12:00). This archetype represents a professionally oriented reader with morning consultation habits.

Beyond structural preferences, the thematic composition of these archetypes also reflects the influence of specific events within the observation period. The concentration on Tennis is linked to Wimbledon 2025 and the coverage of Jannik Sinner’s results, while the weight of Sensitive Topics reflects the editorial prominence of specific news stories during July 2025. This sensitivity to current events is itself a relevant property: it shows that the model captures the actual dynamics of content consumption rather than static demographic categories.

The archetypes also play an operational role in addressing cold start. Figure 3 shows the trajectory of a new user (User 1247) in the Context Space over successive interactions. The user, accessing from a mobile Android device, initially oscillates between multiple archetypes as the first interactions span Sensitive Topics, Sports, Fine Art and Business categories. As more interactions accumulate, the profile stabilizes near the second archetype (the casual mobile reader), demonstrating how the warm-start mechanism provides an immediate, behaviourally grounded placement that is progressively refined.

5. Discussion

The experimental results confirm that the combination of representation learning (FM), clustering and formal context modelling provides a profiling approach that balances predictive quality, computational efficiency and interpretability on real navigation data without persistent identifiers.

The model comparison shows that the FM is not chosen at the expense of accuracy: the differences with xDeepFM are marginal and inconsistent in direction, while the training cost advantage is substantial. In an operational context with continuous editorial drift and the need for periodic retraining, this efficiency gap is what makes the retraining loop feasible at all. The incremental evaluation provides direct evidence of this point: the model converges to a useful representation with a moderate amount of data (peaking at v_3) and does not degrade catastrophically as the distribution shifts, while training time remains proportional and bounded.

The role of CST in this architecture is what distinguishes the approach from a standard embedding-plus-clustering pipeline. The three structural properties introduced in Section 3.3 (dimensional de-

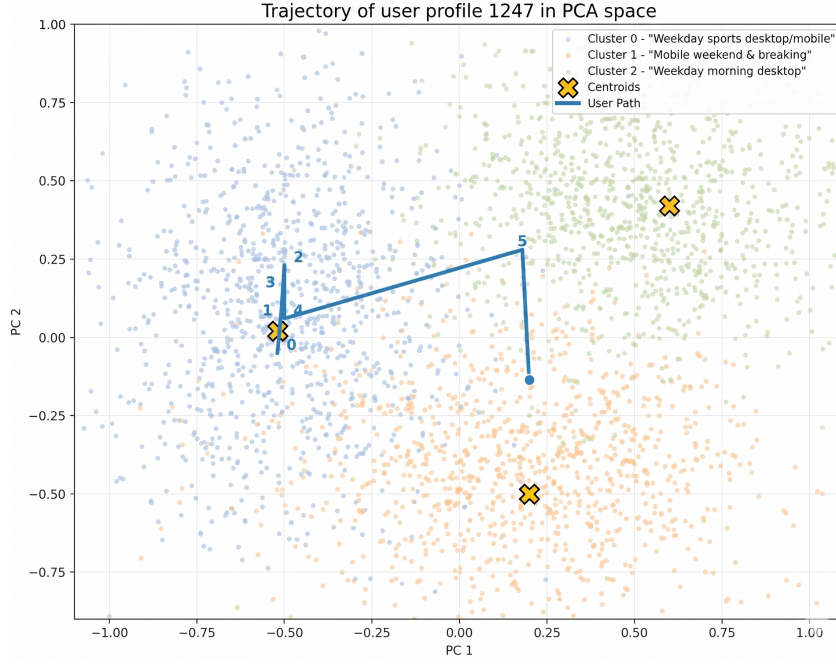


Figure 3: Trajectory of User 1247 in the PCA-projected Context Space. Numbers indicate the temporal order of interactions. The profile migrates from an initial hybrid position toward the second archetype centroid as interactions accumulate.

composition, directional profiles, and stable archetype regions) are confirmed by the experimental evidence to be operationally effective rather than purely theoretical: each cluster centroid admits a direct editorial reading in terms of dominant device, temporal and thematic characteristics, and the warm-start results demonstrate that the geometry of the Context Space supports meaningful placement from the very first interaction.

The cold-start analysis quantifies a structural limitation of profiling without persistent identifiers, which the warm-start mechanism does not close entirely but mitigates by providing a behaviourally grounded initial placement. The trajectory of User 1247 illustrates this concretely: even with a hybrid initial pattern, the system assigns a plausible archetype from the first interaction and refines it as evidence accumulates, suggesting that the Context Space structure imposes a useful inductive bias on the representation.

These properties amount to a form of process transparency built into the representation itself, rather than added as a post-hoc explanation layer: every step (identifier replacement, content labelling, profile construction, archetype assignment) operates on named, inspectable inputs and outputs. This is a direct consequence of the design choices and reflects the operational reality that a system processing behavioural data across hundreds of publisher sites must be able to answer the question “why was this user placed in this segment?” at any time.

6. Conclusion and Future Work

This paper has presented Hagrid, showing that the combination of FM-based representation learning and CST-based context modeling yields a profiling pipeline that is competitive in predictive performance, approximately sixteen times cheaper to train than deeper alternatives, and structurally interpretable for editorial use. The evaluation indicates that the Factorization Machine baseline matches deeper models at a fraction of the cost, while CST integration produces a semantically structured latent space. The proposed archetype-based warm-start mechanism further provides a behaviourally grounded solution to the cold-start problem in this setting.

The analysis also shows that the learned archetypes are sensitive to transient editorial events, such

as Wimbledon 2025 and coverage of Jannik Sinner, suggesting that the model captures both stable consumption patterns and short-term editorial dynamics. Several limitations should be acknowledged. The empirical analysis is restricted to four publisher domains, limiting generalizability. The classification layer is evaluated only indirectly through downstream performance, and its direct accuracy against IAB labels remains untested. In addition, the privacy guarantees correspond to a pseudonymization regime rather than full anonymization, and a formal threat-model analysis is left for future work. Finally, the contribution of the warm-start mechanism is not fully disentangled from content accumulation effects in cold-start scenarios. Future work will explore sentiment-aware extensions of the content representation and dynamic time-decay mechanisms to better separate short-term trends from long-term user preferences in live publishing environments. An explicit objective of the next phase of the project is also a systematic quantitative evaluation of the warm-start mechanism against unprofiled baselines on a larger and more heterogeneous cold-start cohort.

Acknowledgments

This work has been carried out within the Hagrid project (CUP B89J24000480005, Project n. F/360003/00/X75), funded by MIMIT under the “Coesione Italia – PN RIC 2021–2027” program. The authors thank Evolution Group for the development of the data collection infrastructure and the University of Salerno for fruitful discussions on the architectural design.

Declaration on Generative AI

During the preparation of this work, the authors used generative AI tools solely to assist with spelling check.

References

- [1] European Parliament and Council, Regulation (EU) 2016/679 (General Data Protection Regulation), Official Journal of the European Union, L 119/1, 2016.
- [2] European Parliament and Council, Regulation (EU) 2022/2065 on a Single Market for Digital Services (Digital Services Act), Official Journal of the European Union, L 277/1, 2022.
- [3] European Parliament and Council, Regulation (EU) 2024/1689 laying down harmonized rules on artificial intelligence (AI Act), Official Journal of the European Union, 2024.
- [4] S. Munir, S. Siby, U. Iqbal, S. Englehardt, Z. Shafiq, C. Troncoso, CookieGraph: Understanding and detecting first-party tracking cookies, in: Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS), ACM, 2023, pp. 3490–3504. doi:10.1145/3576915.3616586.
- [5] S. Kulkarni, S. F. Rodd, Context aware recommendation systems: A review of the state of the art techniques, Computer Science Review 37 (2020) 100255. doi:10.1016/j.cosrev.2020.100255.
- [6] N. M. Villegas, C. Sánchez, J. Díaz-Cely, G. Tamura, Characterizing context-aware recommender systems: A systematic literature review, Knowledge-Based Systems 140 (2018) 173–200. doi:10.1016/j.knosys.2017.11.003.
- [7] P. Eckersley, How unique is your web browser?, in: Proceedings of the 10th International Conference on Privacy Enhancing Technologies (PETS), Springer, 2010, pp. 1–18. doi:10.1007/978-3-642-14527-8_1.
- [8] P. Laperdrix, N. Bielova, B. Baudry, G. Avoine, Browser fingerprinting: A survey, ACM Transactions on the Web 14 (2020) 1–33. doi:10.1145/3386040.
- [9] IAB Tech Lab, Content Taxonomy 3.0, <https://iabtechlab.com/standards/content-taxonomy/>, 2021. Accessed: 2026-04-26.
- [10] S. Rendle, Factorization machines, in: 2010 IEEE International Conference on Data Mining (ICDM), 2010, pp. 995–1000. doi:10.1109/ICDM.2010.127.

- [11] A. Padovitz, S. W. Loke, A. Zaslavsky, Towards a theory of context spaces, in: Proceedings of the Second IEEE Annual Conference on Pervasive Computing and Communications Workshops (PerCom Workshops), IEEE, 2004, pp. 38–42. doi:10.1109/PERCOMW.2004.1276902.
- [12] G. D’Aniello, M. Gaeta, P. Policastro, Machine learning-based Context Space Theory, in: 2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2023, pp. 3683–3689. doi:10.1109/SMC53992.2023.10394103.
- [13] C. I. Eke, A. A. Norman, L. Shuib, H. F. Nweke, A survey of user profiling: State-of-the-art, challenges, and solutions, IEEE Access 7 (2019) 144907–144924. doi:10.1109/ACCESS.2019.2944243.
- [14] S. Li, H. Zhao, A survey on representation learning for user modeling, in: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI), 2020, pp. 4997–5003. doi:10.24963/ijcai.2020/695.
- [15] G. D’Aniello, M. Della Corte, M. Gaeta, A recommendation system based on fuzzy signature, IEEE Access 14 (2026) 9975–9985. doi:10.1109/ACCESS.2026.3656087.
- [16] Google, The Topics API, <https://privacysandbox.com/proposals/topics>, 2022. Accessed: 2026-04-01.
- [17] Q. Yang, Y. Liu, T. Chen, Y. Tong, Federated machine learning: Concept and applications, ACM Transactions on Intelligent Systems and Technology 10 (2019) 1–19. doi:10.1145/3298981.
- [18] F. Wu, Y. Qiao, J.-H. Chen, C. Wu, T. Qi, J. Lian, D. Liu, X. Xie, J. Gao, W. Wu, M. Zhou, MIND: A large-scale dataset for news recommendation, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), 2020, pp. 3597–3606. doi:10.18653/v1/2020.acl-main.331.
- [19] J. A. Gulla, L. Zhang, P. Liu, Ö. Özgöbek, X. Su, The Adressa dataset for news recommendation, in: Proceedings of the International Conference on Web Intelligence (WI), ACM, 2017, pp. 1042–1048. doi:10.1145/3106426.3109436.
- [20] S. Rendle, Factorization machines with libFM, ACM Transactions on Intelligent Systems and Technology 3 (2012) 1–22. doi:10.1145/2168752.2168771.
- [21] H. Guo, R. Tang, Y. Ye, Z. Li, X. He, DeepFM: A factorization-machine based neural network for CTR prediction, in: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI), 2017, pp. 1725–1731. doi:10.24963/ijcai.2017/239.
- [22] J. Lian, X. Zhou, F. Zhang, Z. Chen, X. Xie, G. Sun, xDeepFM: Combining explicit and implicit feature interactions for recommender systems, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018, pp. 1754–1763. doi:10.1145/3219819.3220023.
- [23] X. He, T.-S. Chua, Neural factorization machines for sparse predictive analytics, in: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2017, pp. 355–364. doi:10.1145/3077136.3080777.
- [24] J. Li, P. Sun, C. Jiang, W. Ma, Q. Ai, M. Zhang, A situation-aware enhancer for personalized recommendation, in: Database Systems for Advanced Applications (DASFAA 2024), volume 14852 of *Lecture Notes in Computer Science*, Springer, 2025, pp. 99–114. doi:10.1007/978-981-97-5555-4_7.
- [25] L. Aliberti, G. D’Aniello, M. Gaeta, D. Perfetto, D. Söfker, Integrating situation awareness into intelligent recommender systems, in: 2026 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA), IEEE, 2026. To appear.
- [26] A. Boytsov, A. Zaslavsky, K. Synnes, Extending context spaces theory by predicting run-time context, in: Smart Spaces and Next Generation Wired/Wireless Networking (NEW2AN/ruSMART 2009), volume 5764 of *Lecture Notes in Computer Science*, Springer, 2009, pp. 8–21. doi:10.1007/978-3-642-04190-7_2.
- [27] F. Colace, G. D’Aniello, M. De Santo, R. Gaeta, G. Zuchtriegel, Situation-aware cyber-physical-social system for cultural heritage, Intelligent Systems with Applications 27 (2025) 200544. doi:10.1016/j.iswa.2025.200544.
- [28] F. Buccafurri, L. Fotia, G. Lax, Privacy-preserving resource evaluation in social networks, in: Proceedings of the 10th Annual International Conference on Privacy, Security and Trust (PST 2012), IEEE, 2012, pp. 51–58. doi:10.1109/PST.2012.6297919.
- [29] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using siamese BERT-networks,

- in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2019, pp. 3982–3992. doi:10.18653/v1/D19-1410.
- [30] J. Johnson, M. Douze, H. Jégou, Billion-scale similarity search with GPUs, *IEEE Transactions on Big Data* 7 (2019) 535–547. doi:10.1109/TBDATA.2019.2921572.
- [31] S. Rendle, C. Freudenthaler, Z. Gantner, L. Schmidt-Thieme, BPR: Bayesian personalized ranking from implicit feedback, in: Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence (UAI), 2009, pp. 452–461.
- [32] C. D. Manning, P. Raghavan, H. Schütze, *Introduction to Information Retrieval*, Cambridge University Press, 2008.
- [33] K. Järvelin, J. Kekäläinen, Cumulated gain-based evaluation of IR techniques, *ACM Transactions on Information Systems* 20 (2002) 422–446. doi:10.1145/582415.582418.
- [34] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, T.-S. Chua, Neural collaborative filtering, in: Proceedings of the 26th International Conference on World Wide Web (WWW), 2017, pp. 173–182. doi:10.1145/3038912.3052569.