Desarrollo de un Modelo de Recomendación Musical Contextualizado aplicando redes trasnformers

QUICAÑO MIRANDA, Victor Alejandro

Orientador: Prof Dr./Mag./Ing. Nombre del Asesor

Plan de Tesis presentado la Escuela Profesional Ciencia de la Computación como paso previo a la elaboración de la Tesis Profesional.

UNSA - Universidad Nacional de San Agustín de Arequipa
Junio de 2025

Abreviaturas

Índice

Índice de cuadros

Índice de figuras

1. Introducción

La música, con su capacidad para evocar emociones profundas y universales, ha sido durante siglos un medio esencial de expresión humana. Las canciones, en particular, combinan letra y melodía para crear experiencias emocionales únicas, capaces de influir en el estado de ánimo y los sentimientos de los oyentes. En la era digital, los sistemas de recomendación musical han buscado emular esta experiencia personalizada, sin embargo, su efectividad se ve limitada por el uso intensivo de datos sensibles y la falta de un modelado contextual profundo [5, 6].

Los enfoques tradicionales de recomendación musical suelen basarse en historial de reproducciones y patrones de consumo previos, lo cual requiere centralizar grandes volúmenes de datos de usuario y plantea desafíos críticos de privacidad, cumplimiento de normativas como el GDPR y riesgos de seguridad de la información. Además, al no considerar variables contextuales —como clima, ubicación, estado emocional, compañía o actividad realizada— estos sistemas no logran capturar la dinámica real de las preferencias musicales, limitando su capacidad de ofrecer recomendaciones verdaderamente relevantes y situacionales [2, 7].

Recientemente, las arquitecturas basadas en Transformers han demostrado un gran potencial para modelar datos secuenciales y capturar dependencias a largo plazo en tareas de procesamiento de lenguaje y visión por computador. En el contexto musical, los Transformers permiten integrar información multimodal y contextual de forma eficiente, ofreciendo representaciones ricas que pueden adaptarse a cambios en tiempo real en el estado del usuario [7]. De igual manera, el aprendizaje federado surge como una solución para preservar la privacidad, al mantener los datos sensoriales y contextuales en el dispositivo del usuario y solo compartir actualizaciones de modelo agregadas [2].

En este trabajo proponemos el diseño y desarrollo de un sistema de recomendación musical contextualizado que combina—por primera vez—una arquitectura Transformer con un esquema de aprendizaje federado. Nuestra contribución principal radica en:

1. Emplear un Transformer adaptado para incorporar factores contextuales dinámicos (clima, estado anímico, actividad) en la representación de preferencias musicales.
2. Integrar un protocolo de aprendizaje federado que garantice la privacidad y seguridad de los datos locales del usuario.
3. Evaluar el rendimiento de nuestro modelo en escenarios reales de uso, comparándolo con sistemas centralizados y no contextuales.

Con ello, aspiramos a avanzar en la personalización musical, ofreciendo recomendaciones precisas y emocionalmente relevantes, sin comprometer la privacidad del usuario ni depender de grandes repositorios de datos centralizados.

2. Trabajos relacionados

2.1. An actor-critic based recommender system with context-aware user modeling [1]

El artículo An actor-critic based recommender system with context-aware user modeling presenta un enfoque de Deep Reinforcement Learning (DRL) que modela el problema de recomendación como un Proceso de Decisión de Markov, donde el agente —dividido en un actor y un crítico— genera recomendaciones teniendo en cuenta el contexto dinámico del usuario. Propone dos estrategias de modelado de estado contextual: Context-based Zero Weighting (CsZW), que filtra por completo el histórico irrelevante, y Context-based Attention Weighting (CsAW), que aplica atención para mantener diversidad. Asimismo, introduce un agente de recomendación lista-por-lista que optimiza con retroalimentación acumulada. Los experimentos, realizados sobre los conjuntos DePaul Movie y LDos-Comoda, muestran mejoras de hasta un 8

La metodología actor-critic contextual de este artículo aporta al desarrollo de un recomendador musical contextualizado dos ideas esenciales. Primero, demuestra la viabilidad de integrar información de contexto en el proceso de recomendación mediante modelado de estado adaptativo, lo cual sienta las bases para incorporar factores externos —como clima o estado de ánimo— en las representaciones del usuario. Segundo, la adopción de un agente lista-por-lista ilustra cómo recolectar y aprovechar retroalimentación secuencial, estrategia que puede trasladarse al dominio musical para refinar recomendaciones en tiempo real. Estas aportaciones inspiran el diseño de nuestro sistema Transformer-federado, al combinar el modelado contextual fino del DRL actor-critic con la capacidad de los Transformers para capturar dependencias a largo plazo y el aprendizaje federado para preservar la privacidad del usuario.

2.2. Self-Attention Mechanism-Based Federated Learning Model for Cross Context Recommendation System [2]

El artículo Self-Attention Mechanism-Based Federated Learning Model for Cross Context Recommendation System propone un modelo híbrido denominado SAFL que combina autoatención y aprendizaje federado para abordar la recomendación entre contextos. El componente de autoatención extrae y pondera dinámicamente las preferencias del usuario a partir de múltiples dominios contextuales, mientras que el esquema federado permite entrenar de forma colaborativa sin compartir datos sensibles. De esta manera, SAFL mitiga los problemas de escasez de datos y cold-start al transferir conocimiento entre contextos, y garantiza la privacidad al mantener la información del usuario alojada localmente en cada dispositivo.

La arquitectura SAFL aporta al desarrollo de nuestro recomendador musical contextualizado dos elementos clave. Primero, demuestra cómo la autoatención puede integrarse para modelar preferencias contextuales complejas —por ejemplo, adaptando la recomen-

dación musical a factores como ambiente, actividad o estado de ánimo— aprovechando la capacidad de los Transformers para capturar relaciones secuenciales y contextuales. Segundo, muestra la viabilidad del aprendizaje federado para preservar la privacidad del usuario, un requisito esencial cuando se manejan datos sensibles de hábitos de escucha y contexto personal. Estas ideas se incorporarán en nuestro sistema Transformer-federado para lograr recomendaciones musicales contextualizadas, precisas y seguras.

2.3. Aggregating Contextual Information for Multi-Criteria Online Music Recommendations [3]

El artículo presenta CAMCMusic un aporte significativo en el ámbito de los sistemas de recomendación musical, particularmente cuando se analiza desde la perspectiva del aprendizaje federado y la personalización contextual. En un entorno donde las preferencias musicales están profundamente influenciadas por factores subjetivos y variables contextuales, CAMCMusic ofrece un enfoque innovador al integrar la toma de decisiones multicriterio (MCDM) con una conciencia contextual dinámica. Esta propuesta representa un paso importante hacia la superación de las limitaciones de los sistemas tradicionales que, por lo general, dependen de calificaciones explícitas, atributos fijos del usuario o características estáticas de las canciones. La capacidad de CAMCMusic para generar recomendaciones relevantes sin necesidad de información directa del usuario abre nuevas posibilidades para desarrollar sistemas de recomendación más privados y escalables.

Desde la óptica del aprendizaje federado, CAMCMusic se alinea con los principios de minimizar la necesidad de recopilar información sensible directamente del usuario. Aunque el artículo no implementa explícitamente un enfoque federado, su estrategia de utilizar datos preexistentes y recomendaciones estereotipadas se puede complementar eficazmente con técnicas de aprendizaje federado para proteger la privacidad del usuario. Al combinar estos dos enfoques, se podría fortalecer la personalización musical respetando los principios de privacidad diferencial y aprendizaje descentralizado. Este vínculo es especialmente relevante en un contexto donde la recopilación de datos emocionales, de ubicación o estado físico puede presentar riesgos éticos y legales si no se maneja cuidadosamente.

Además, la relevancia de CAMCMusic se magnifica al abordar el problema del .ªrranque en frío", uno de los desafíos persistentes en los sistemas de recomendación. Al evitar depender exclusivamente de información previa del usuario, y al construir recomendaciones basadas en patrones de comportamiento contextual agrupados, el sistema no solo mejora la calidad de las recomendaciones, sino que también facilita la adopción en escenarios con datos limitados o usuarios nuevos. Esta capacidad puede aprovecharse dentro de entornos de aprendizaje federado para hacer recomendaciones útiles sin compartir datos personales entre dispositivos o servidores.

Finalmente, CAMCMusic sienta las bases para el desarrollo de futuros sistemas de recomendación que integren el aprendizaje federado con una estructura multicriterio contextual. Esta combinación permitiría una evolución significativa en la personalización musical: sistemas más seguros, adaptativos y éticamente sostenibles que comprenden mejor

las complejidades del comportamiento humano sin comprometer la privacidad. Por lo tanto, el artículo no solo contribuye al estado del arte en recomendación musical, sino que también abre caminos hacia una integración más profunda entre inteligencia artificial, privacidad y experiencia del usuario.

2.4. Listener Modeling and Context-Aware Music Recommendation Based on Country Archetypes [4]

El artículo "Listener Modeling and Context-Aware Music Recommendation Based on Country Archetypes" representa una contribución destacada al campo de los sistemas de recomendación musical al incorporar de manera efectiva el contexto geográfico del usuario en modelos avanzados de aprendizaje profundo. Su enfoque innovador parte del análisis de grandes volúmenes de datos de comportamiento musical y propone una arquitectura de autoencoder variacional (VAE) mejorada que integra información contextual sin depender de fuentes externas. Esta estrategia no solo optimiza la personalización de las recomendaciones, sino que también mantiene un enfoque auto-sostenible, lo cual resulta especialmente valioso para plataformas que buscan soluciones escalables y eficientes sin requerir datos sensibles o difíciles de obtener.

Desde la perspectiva del aprendizaje federado, la propuesta del artículo abre caminos para el desarrollo de sistemas descentralizados y respetuosos con la privacidad. Aunque el modelo actual no aplica directamente técnicas federadas, su dependencia exclusiva del historial de escucha y del país autoinformado por el usuario facilita una posible implementación futura en entornos federados. En tales escenarios, los modelos podrían entrenarse localmente en los dispositivos de los usuarios, manteniendo su información personal segura mientras se sigue aprovechando el contexto geográfico para mejorar la calidad de las recomendaciones. Esto es especialmente relevante dado el creciente énfasis en la privacidad de los datos en la inteligencia artificial moderna.

Además, la investigación subraya un aspecto que a menudo es pasado por alto: la importancia del contexto cultural y geográfico en las preferencias musicales. A través de la identificación de arquetipos de países mediante técnicas no supervisadas, los autores logran capturar patrones colectivos de escucha sin imponer supuestos culturales explícitos. Esta abstracción de alto nivel es particularmente útil para construir modelos más generalizables que pueden adaptarse a distintos grupos de usuarios sin requerir personalización individual exhaustiva. La integración de estos arquetipos en una arquitectura VAE con mecanismos de gating permite que las preferencias individuales sean moduladas por el entorno colectivo, reflejando una forma sofisticada de conciencia contextual.

En conjunto, el artículo no solo ofrece un avance técnico en términos de modelado de usuario consciente del contexto, sino que también propone una estructura sólida que puede evolucionar hacia entornos de aprendizaje federado. En el contexto más amplio del desarrollo de sistemas de recomendación éticos, escalables y personalizados, esta investigación proporciona herramientas y perspectivas clave para diseñar tecnologías que equilibran efectividad, adaptabilidad y privacidad. La combinación de geolocalización auto-informada,

análisis de comportamiento y aprendizaje profundo representa una fórmula poderosa para futuras soluciones en la recomendación musical moderna.

3. Propuesta

3.1. Construcción del Dataset Contextualizado

La primera etapa consiste en la elaboración de un dataset que integre información de preferencias musicales junto con datos contextuales:

- Se toma como base el dataset **Music4All-Onion**[8], [9], el cual contiene información detallada sobre interacciones usuario-canción, así como múltiples representaciones musicales (MFCC, VAD, géneros, letras, etc.).
- A este conjunto se le incorpora información contextual proveniente del dataset **ExtraSensory**[10], el cual incluye datos capturados desde sensores de dispositivos móviles y wearables (por ejemplo, actividad física, localización, sonido ambiente, uso del teléfono, etc.).
- La unión de ambos conjuntos se realiza de forma simulada, asociando los usuarios con mayor número y diversidad de registros en ambas fuentes. De este modo, se simulan escenarios donde las decisiones de escucha se ven influenciadas por el contexto ambiental y conductual del usuario.

El resultado es un conjunto de datos enriquecido que representa de manera aproximada cómo varía el comportamiento musical del usuario en diferentes situaciones reales.

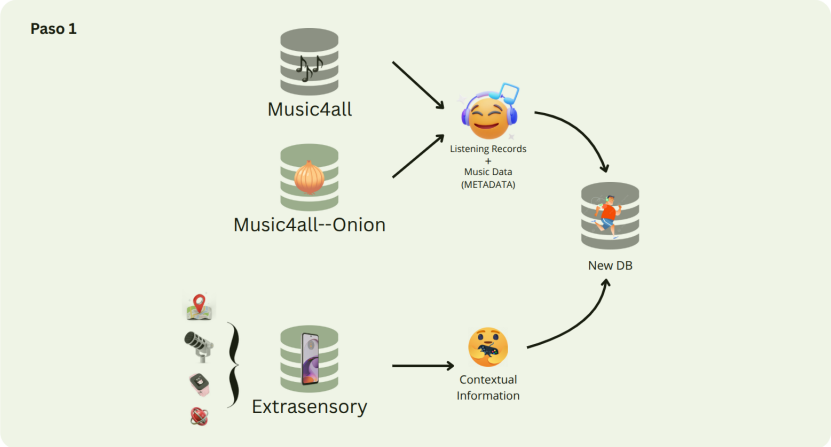

Figura 1: Pipeline del modelo de base de datos.

3.2. Validación del Dataset con un Modelo Transformer

Una vez construido el dataset contextualizado, se procede a la validación de los registros a través de un modelo basado en arquitecturas **Transformer**. Esta fase tiene como objetivo comprobar que las canciones asociadas a cada contexto son coherentes con los patrones de gusto del usuario:

- El modelo Transformer es entrenado para predecir canciones que un usuario tendería a escuchar bajo un determinado contexto.
- Se evalúa si las canciones asignadas en el dataset simulado coinciden con las predicciones generadas por el modelo, lo cual sirve como verificación de consistencia.
- Esta etapa también permite ajustar los pesos del modelo base y definir una representación inicial sólida para ser utilizada posteriormente en el entorno federado.

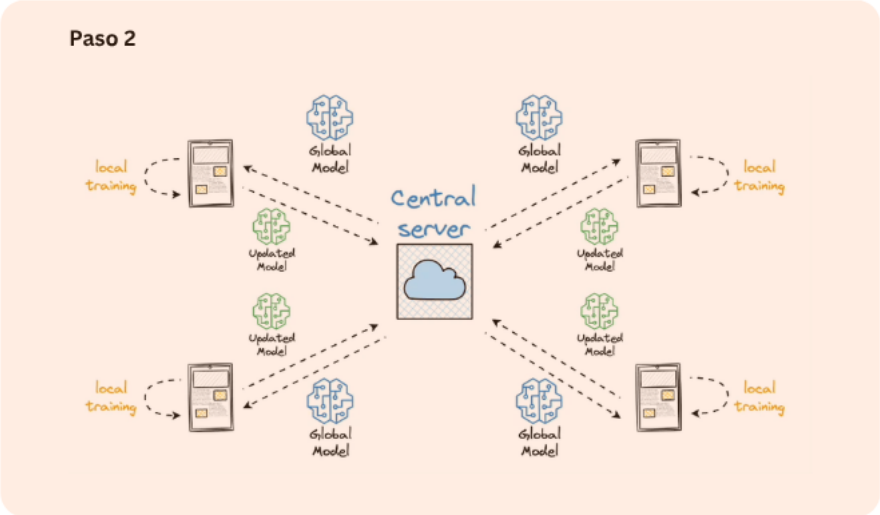

Figura 2: Modelo de Aprendizaje Federado.

Referencias

[1] M. Bukhari, M. Maqsood, and F. Adil, "An actor-critic based recommender system with context-aware user modeling," *Artif. Intell. Rev.*, vol. 58, no. 5, Feb. 2025.

[2] N. K. Singh, D. S. Tomar, M. Shabaz, I. Keshta, M. Soni, D. R. Sahu, M. S. Bhende, A. K. Nandanwar, and G. Vishwakarma, "Self-attention mechanism-based federated learning model for cross context recommendation system," *IEEE Transactions on Consumer Electronics*, vol. 70, no. 1, p. 2687–2695, Feb. 2024.

[3] J. Liu, "Aggregating contextual information for multi-criteria online music recommendations," *IEEE Access*, vol. 13, pp. 8790–8805, 2025.

[4] M. Schedl, C. Bauer, W. Reisinger, D. Kowald, and E. Lex, "Listener modeling and context-aware music recommendation based on country archetypes," *Front. Artif. Intell.*, vol. 3, p. 508725, 2020.

[5] B. O. Ghanshyambhai, "Context-aware music embedding in silent videos leveraging transformer architectures: A review," *International Journal of Science, Engineering and Technology*, vol. 13, no. 1, pp. 1–27, Jan. 2025.

[6] C. Huang, T. Yu, K. Xie, S. Zhang, L. Yao, and J. McAuley, "Foundation models for recommender systems: A survey and new perspectives," 2024.

[7] O. Badhe, D. K. Sutaria, and D. V. Shorthiya, "Context-aware music embedding in silent videos leveraging transformer architectures: A review." [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract$_id = 5108046$

[8] I. A. Pegoraro Santana, F. Pinhelli, J. Donini, L. Catharin, R. B. Mangolin, Y. M. e. G. da Costa, V. Delisandra Feltrim, and M. A. Domingues, "Music4all: A new music database and its applications," in *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*. IEEE, Jul. 2020, p. 399–404. [Online]. Available: http://dx.doi.org/10.1109/IWSSIP48289.2020.9145170

[9] M. Moscati, E. Parada-Cabaleiro, Y. Deldjoo, E. Zangerle, and M. Schedl, "Music4all-onion – a large-scale multi-faceted content-centric music recommendation dataset," in *Proceedings of the 31st ACM International Conference on Information amp; Knowledge Management*, ser. CIKM '22. ACM, Oct. 2022, p. 4339–4343. [Online]. Available: http://dx.doi.org/10.1145/3511808.3557656

[10] Y. Vaizman, K. Ellis, G. Lanckriet, and N. Weibel, "Extrasensory app: Data collection in-the-wild with rich user interface to self-report behavior," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, ser. CHI '18. ACM, Apr. 2018, p. 1–12. [Online]. Available: http://dx.doi.org/10.1145/3173574.3174128