Universidade Federal de São Carlos– UFSCar
Centro de Ciências Exatas e de Tecnologia– CCET
Departamento de Computação– DC
Programa de Pós-Graduação em Ciência da Computação– PPGCC

Pietro Lo Presti Campos

Avaliação de Bandits Contextuais para Recomendação: Temporalidade e Limitações

São Carlos
2025

Pietro Lo Presti Campos

Avaliação de Bandits Contextuais para Recomendação: Temporalidade e Limitações

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Centro de Ciências Exatas e de Tecnologia da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Área de concentração: Metodologias e Técnicas de Computação

Orientador: Prof. Dr. Tiago Agostinho de Almeida

São Carlos
2025

Campos, Pietro Lo Presti

Avaliação de bandits contextuais para recomendação: temporalidade e limitações / Pietro Lo Presti Campos -- 2025.
142f.

Dissertação (Mestrado) - Universidade Federal de São Carlos, campus São Carlos, São Carlos
Orientador (a): Tiago Agostinho de Almeida
Banca Examinadora: Tiago Agostinho de Almeida, Alan Demétrius Baria Valejo, Marcelo Garcia Manzato
Bibliografia

1. Sistemas de recomendação. 2. Modelagem temporal. 3. Bandits contextuais. I. Campos, Pietro Lo Presti. II. Título.

Ficha catalográfica desenvolvida pela Secretaria Geral de Informática (SIn)

DADOS FORNECIDOS PELO AUTOR

Bibliotecário responsável: Arildo Martins - CRB/8 7180

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Exatas e de Tecnologia
Programa de Pós-Graduação em Ciência da Computação

Folha de Aprovação

Defesa de Dissertação de Mestrado do candidato Pietro Lo Presti Campos, realizada em 01/10/2025.

Comissão Julgadora:

Prof. Dr. Tiago Agostinho de Almeida (UFSCar)

Prof. Dr. Alan Demétrius Baria Valejo (UFSCar)

Prof. Dr. Marcelo Garcia Manzato (ICMC/USP)

O Relatório de Defesa assinado pelos membros da Comissão Julgadora encontra-se arquivado junto ao Programa de Pós-Graduação em Ciência da Computação.

Dedico este trabalho a todos que abriram meu caminho
e àqueles que nele encontrarão novos passos.

Agradecimentos

Sou grato à minha família, que desde o início da minha vida escolar e acadêmica esteve presente, oferecendo incentivo e condições para que eu pudesse chegar até aqui.

Agradeço à minha esposa Júlia, pela parceria, carinho e apoio ao longo desta caminhada.

Ao grupo de pesquisa RecSys da UFSCar, pela participação em todas as etapas deste trabalho. Em especial Pedro, Gregório e Rafael, pelo apoio constante, pelas discussões e pelas contribuições que enriqueceram este projeto. Foi uma experiência muito valiosa mantermos o projeto ativo de forma contínua ao longo do tempo, sempre com trocas de conhecimento regulares e construtivas. Tenho orgulho do trabalho desenvolvido pelo grupo e da trajetória que construímos juntos.

Ao meu orientador, professor Tiago, pela oportunidade única de desenvolver este trabalho e pela confiança depositada.

Por fim, à UFSCar, pelo ambiente acadêmico de qualidade.

"Tens o direito à ação, mas nunca aos frutos dela.
Jamais busques o fruto da ação; e nunca deixes de agir."
(Bhagavad Gita, II.47)

Resumo

Em um cenário digital em que usuários são expostos diariamente a um volume massivo de conteúdos, os sistemas de recomendação desempenham um papel essencial ao filtrar e personalizar informações. Esses sistemas, no entanto, enfrentam o dilema clássico entre exploração (apresentar novos itens) e aprofundamento (reforçar preferências já conhecidas). Encontrar o equilíbrio ideal entre esses dois comportamentos é um dos maiores desafios da área, especialmente em abordagens adaptativas como os *Multi-Armed Bandits* (MAB) contextuais, que aprendem continuamente a partir das interações de cada usuário.

Este trabalho teve início com a investigação de diferentes algoritmos lineares de MAB em cenários de recomendação e avaliação *offline*. Durante esses experimentos, observou-se um viés sistemático nas métricas tradicionais, que favoreciam métodos puramente gulosos (sem exploração), comprometendo a análise de estratégias exploratórias e a comparação justa entre políticas.

Para contornar essas limitações, foi proposta e implementada uma nova metodologia de avaliação *online* em ambiente simulado. O simulador *KuaiSim*, baseado na base de dados *KuaiRand*, foi extensivamente adaptado para suportar interações multi-sessão, modelagem contextual e dependência temporal. Essa infraestrutura possibilitou investigar, de forma mais realista, como fatores temporais influenciam o comportamento de recomendação.

Com base nesse novo ambiente, foi desenvolvido o método temporal *Time-Aware Lin-Boltzmann*, que combina modelos lineares com exploração *Boltzmann* e ajusta dinamicamente o parâmetro de temperatura de acordo com o intervalo de tempo entre interações. A intuição é que usuários que retornam rapidamente tendem a preferir recomendações mais alinhadas a seus interesses prévios, enquanto retornos mais longos podem indicar maior disposição para explorar novas opções.

Os experimentos realizados no simulador indicam que o uso do tempo melhora métricas de diversidade e cobertura em relação aos métodos de referência, sugerindo que a informação temporal é um sinal relevante para modular o equilíbrio entre exploração e aprofundamento. Esses resultados abrem novas perspectivas para sistemas de recomen-

dação sensíveis ao comportamento temporal do usuário.

As principais contribuições deste trabalho incluem: (i) a identificação e análise do viés em protocolos *offline* de avaliação de MAB lineares; (ii) o desenvolvimento de uma estrutura de simulação *online* baseada em sessões temporais; (iii) a proposição do algoritmo *Time-Aware LinBoltzmann*, que introduz consciência temporal no processo de exploração; e (iv) a discussão de implicações, limitações e oportunidades futuras para o uso de informação temporal em sistemas de recomendação.

Palavras-chave: Sistemas de recomendação. Multi-Armed Bandits. Consciência Temporal. Avaliação Online.

Abstract

In a digital environment where users are daily exposed to a massive volume of content, recommender systems play an essential role in filtering and personalizing information. These systems, however, face the classic dilemma between exploration (introducing new items) and exploitation (reinforcing known preferences). Finding the ideal balance between these two behaviors remains one of the major challenges in the field, especially in adaptive approaches such as contextual *Multi-Armed Bandits* (MAB), which learn continuously from user interactions over time.

This work began with the investigation of different linear MAB algorithms in recommendation and offline evaluation scenarios. During these experiments, a systematic bias was observed in traditional metrics, favoring purely greedy methods (without exploration) and compromising both the analysis of exploratory strategies and the fair comparison between policies.

To overcome these limitations, a new online evaluation methodology was proposed and implemented in a simulated environment. The *KuaiSim* simulator, based on the *KuaiRand* dataset, was extensively adapted to support multi-session interactions, contextual modeling, and temporal dependency. This infrastructure enabled a more realistic investigation of how temporal factors influence recommendation behavior.

Building upon this new environment, the temporal method *Time-Aware LinBoltzmann* was developed. It combines linear models with *Boltzmann* exploration and dynamically adjusts the temperature parameter according to the time interval between user interactions. The underlying intuition is that users who return quickly tend to prefer recommendations aligned with their previous interests, while longer return intervals may indicate a greater willingness to explore new options.

Experiments conducted in the simulator show that incorporating temporal information improves diversity and coverage metrics compared to baseline methods, suggesting that time is a relevant signal for modulating the balance between exploration and exploitation. These findings open new perspectives for recommender systems that are sensitive to users'

temporal behavior.

The main contributions of this work include: (i) the identification and analysis of bias in offline evaluation protocols for linear MABs; (ii) the development of an online simulation framework based on temporal sessions; (iii) the proposal of the *Time-Aware LinBoltzmann* algorithm, which introduces temporal awareness into the exploration process; and (iv) the discussion of implications, limitations, and future opportunities for incorporating temporal information into recommender systems.

Keywords: Recommender Systems. Multi-Armed Bandits. Temporal Awareness. Online Evaluation.

Lista de ilustrações

Lista de tabelas

Lista de siglas

A/B *A/B Testing*
(Testes A/B)

AdaLinUCB *Adaptive Linear Upper Confidence Bound*
(UCB Linear Adaptativo)

ALS *Alternating Least Squares*
(Mínimos Quadrados Alternados)

AUC *Area Under the Curve*
(Área Sob a Curva)

CMAB *Contextual Multi-Armed Bandits*
(Bandidos de Múltiplos Braços Contextuais)

CLUCB2 *Conservative Linear UCB (v2)*
Variante conservadora do UCB Linear

CTR *Click-Through Rate*
(Taxa de Cliques)

DM *Direct Method*
(Método Direto)

DR *Doubly Robust*

DNN *Deep Neural Network*
(Rede Neural Profunda)

ILD *Intra-List Diversity*
(Diversidade Intra-lista)

IPW *Inverse Propensity Weighting*

Lin *Linear*
(baseline determinístico)

LinGreedy *ε-greedy linear*
(Método Guloso-ε Linear)

LinTS *Linear Thompson Sampling*

LinUCB *Linear Upper Confidence Bounds*

LNUCB-TA *Time-Aware Linear UCB*
(UCB Linear Sensível ao Tempo)

MAB *Multi-Armed Bandits*
(Bandidos de Múltiplos Braços)

NDCG *Normalized Discounted Cumulative Gain*

OBP *Open Bandit Pipeline*

OPE *Off-Policy Evaluation*

RecSim *Recommender Simulation*
(Simulador de Recomendação)

RecSimNG *Recommender Simulation Next Generation*

RecoGym Ambiente de simulação *RecoGym*

RL *Reinforcement Learning*
(Aprendizado por Reforço)

RL4RS *Reinforcement Learning for Recommender Systems*

RQ *Research Question*
(Pergunta de Pesquisa)

TARS *Time-Aware Recommender Systems*
(Sistemas de Recomendação Sensíveis ao Tempo)

TALB *Time-Aware LinBoltzmann*
(Forma abreviada do Algoritmo LinBoltzmann Sensível ao Tempo)

TS *Thompson Sampling*

UCB *Upper Confidence Bound*

Sumário

Capítulo 1

Introdução

O crescimento contínuo do conteúdo digital aumentou a quantidade de informação disponível e consolidou os sistemas de recomendação como uma infraestrutura essencial para mediar a interação entre usuários e itens. De modo geral, a literatura apresenta três grandes abordagens. A recomendação baseada em conteúdo (*content-based*) utiliza as características do item, como o gênero de um filme ou as palavras-chave de um artigo, para sugerir alternativas semelhantes ao que o usuário já consumiu. A filtragem colaborativa (*collaborative filtering*) considera o comportamento coletivo, ou seja, recomenda itens a partir de padrões compartilhados entre usuários com preferências semelhantes. Já a abordagem híbrida (*hybrid*) combina elementos dos dois métodos anteriores e busca aproveitar as vantagens de cada um.

Apesar dos avanços, algumas limitações estruturais continuam desafiando a área. A esparsidade (*sparsity*) acontece quando existem poucas interações registradas em relação ao grande volume de itens disponíveis, o que dificulta aprender preferências confiáveis. O clássico problema de *cold-start* (início a frio) surge quando não há informações suficientes sobre novos usuários ou novos itens, tornando difícil gerar recomendações relevantes. Já a superespecialização (*overspecialization*) ocorre quando o sistema insiste em recomendar apenas itens muito semelhantes ao histórico do usuário, reduzindo a diversidade das sugestões.

Nos últimos anos, ganhou força a ideia de incorporar o contexto, como o tempo, a localização ou as circunstâncias de consumo, e também de avaliar os sistemas de recomendação com métricas que vão além da acurácia. Diversidade, novidade e cobertura, por exemplo, passaram a ser valorizadas por refletirem melhor a utilidade prática e o impacto de longo prazo de um sistema de recomendação (RESNICK; VARIAN, 1997; ADOMAVICIUS; TUZHILIN, 2005; RICCI; ROKACH; SHAPIRA, 2015; BOBADILLA

et al., 2013).

Em ambientes dinâmicos, como notícias, transmissões *online* (*streaming*) e conteúdo em redes sociais, o dilema entre exploração (*exploration*) e aprofundamento (*exploitation*) torna-se central. Explorar significa dedicar parte das recomendações a itens menos conhecidos, com o objetivo de descobrir preferências ainda não observadas e ampliar a cobertura do catálogo. Já aprofundar significa concentrar as recomendações em itens semelhantes ao histórico do usuário, aproveitando o conhecimento acumulado para maximizar a recompensa imediata.

Esse dilema é clássico nos problemas conhecidos como *Multi-Armed Bandits* (Bandidos de Múltiplos Braços), abreviados como MAB. A metáfora vem das máquinas caça-níqueis: cada "braço" representa uma ação possível, com uma recompensa incerta, e o desafio do agente é descobrir qual braço concede a maior recompensa a longo prazo, escolhendo-o com maior frequência. Em aprendizado por reforço (*reinforcement learning*) e em MAB, diversas estratégias foram propostas para equilibrar exploração e aprofundamento. Entre elas, a política *epsilon-greedy*, que introduz exploração aleatória em parte das decisões; os índices de confiança superior (*Upper Confidence Bounds*, UCB), que priorizam ações com maior incerteza; a exploração proporcional à preferência, como nos métodos *softmax* ou Boltzmann; e a amostragem bayesiana (*Thompson Sampling*), que escolhe ações de acordo com distribuições de probabilidade sobre as recompensas (SUTTON; BARTO, 1998; AUER; CESA-BIANCHI; FISCHER, 2002; LATTIMORE; SZEPESVÁRI, 2020).

No contexto de recomendação personalizada, uma extensão natural são os *bandits* contextuais (*Contextual Multi-Armed Bandits*, CMAB), que incorporam atributos de usuários e itens no processo de decisão. Entre os métodos existentes, as variantes lineares se destacam pela eficiência e escalabilidade, como Lin, LinGreedy, LinUCB e LinTS (LI et al., 2010).

Dentro desse cenário, um aspecto particularmente relevante é a dimensão temporal, a qual não deve ser vista apenas como mais um atributo, mas como um fator que altera a própria lógica de decisão. Estudos em sistemas sensíveis ao tempo (do inglês, *Time-Aware Recommender Systems*, TARS) mostram que preferências variam em múltiplas escalas, do curto ao longo prazo, e sofrem efeitos de recência, esquecimento e sazonalidade (CAMPOS; DÍEZ; CANTADOR, 2014; BOGINA et al., 2023). Além de características de tempo absoluto, como o dia da semana ou a "idade" do item, ganha destaque o tempo relativo entre interações. Esses intervalos, isto é, o tempo entre um consumo e outro, influenciam tanto a escolha das recomendações quanto a probabilidade de retorno do usuário e sua disposição para receber novidades (YE et al., 2020). Ignorar essa dimensão pode comprometer a estimativa de recompensas e reduzir a eficácia da exploração.

Com base nisso, *esta dissertação parte da hipótese de que o intervalo de tempo entre interações do usuário pode calibrar, de forma simples e transparente, o grau de exploração em CMAB.* A intuição é direta: retornos rápidos, quando o usuário volta em pouco tempo,

sugerem engajamento contínuo e favorecem recomendações mais conservadoras, próximas ao histórico. Já retornos demorados podem indicar mudanças de interesse, incentivando uma postura mais exploratória. Assim, em vez de tratar o tempo apenas como parte do vetor de contexto, ele é utilizado como modulador da exploração no próprio mecanismo de seleção de ações. Essa perspectiva é compatível com trabalhos que distinguem padrões temporais absolutos e relativos em recomendações sequenciais (YE et al., 2020; CAMPOS; DÍEZ; CANTADOR, 2014; BOGINA et al., 2023).

No entanto, uma limitação prática precisa ser reconhecida: avaliações *offline* podem favorecer políticas mais gulosas (com pouca, ou nenhuma, exploração), em detrimento de estratégias que buscam explorar novos conteúdos. Isso cria um viés que dificulta estimar de forma justa o impacto de políticas exploratórias.

Dessa forma, três pontos principais motivam este estudo. Primeiro, políticas exploratórias são essenciais para ampliar a diversidade e a cobertura das recomendações, mas correm o risco de serem subestimadas em protocolos *offline*. Segundo, a temporalidade, em especial os intervalos entre consumo, influenciam diretamente o comportamento do usuário e a eficácia da exploração. Terceiro, políticas simples, transparentes e sensíveis ao tempo (*time-aware*) podem facilitar a interpretação dos resultados e orientar o desenvolvimento de sistemas de recomendação mais realistas e eficazes.

1.1 Lacunas de Pesquisa

Com base na revisão da literatura, as principais lacunas de pesquisa podem ser resumidas da seguinte forma:

L1. Limitações da avaliação *offline* em políticas exploratórias: métodos tradicionais partem de suposições (como todas as ações terem chance de ser escolhidas) que raramente se confirmam na prática, levando a avaliações distorcidas.

L2. Viés de seleção dinâmico: a probabilidade de um item ser mostrado e a forma como os usuários escolhem mudam ao longo do tempo (por exemplo, itens novos atraem mais atenção que itens antigos), o que torna insuficientes correções estáticas de viés.

L3. Protocolos temporais limitados: ainda faltam métodos de avaliação que considerem de forma explícita os efeitos do tempo, por exemplo, como as recomendações influenciam a frequência de retorno dos usuários e como medir o envelhecimento das preferências ao longo das sessões.

L4. Desalinhamento de métricas: ainda há foco quase exclusivo em métricas de acurácia ou clique (CTR), com pouca atenção a aspectos como cobertura e diversidade das recomendações.

L5. Pouca exploração de algoritmos sensíveis ao tempo: há escassez de estudos que usam informações temporais (como o intervalo entre consumos) para ajustar dinamicamente o nível de exploração, indo além de tratar o tempo apenas como mais uma variável de contexto.

1.2 Perguntas de Pesquisa

Derivadas das lacunas acima, as perguntas que orientam a investigação teórica e empírica são:

PP1: Como avaliar de maneira justa CMAB com exploração adaptativa? Quais condições e protocolos são mais adequados.

PP2: Como modelar o tempo em CMAB de forma realista? Em especial, como representar o **tempo de retorno** quando ele também é influenciado pelas próprias recomendações.

PP3: Qual o efeito do tempo como fator de exploração? Impactos de usar informações temporais na exploração sobre, por exemplo, recompensa, diversidade e cobertura.

PP4: Quais limitações e desafios práticos dessa abordagem? Restrições de dados, protocolos e generalização.

Essas questões guiam o desenho experimental e a escolha de métricas, e são respondidas ao longo deste trabalho através dos experimentos *offline*, simulações *online* e adaptação de métodos tradicionais de recomendação para incorporação do tempo.

1.3 Hipóteses

Com base nas lacunas metodológicas e nas PPs, foram levantadas as seguintes hipóteses:

H1: Avaliação: Protocolos tradicionais de avaliação *offline* podem subestimar CMAB com exploração adaptativa sob *logs* determinísticos e suporte limitado; avaliações *online* e simulação devem refletir melhor os efeitos reais da exploração.

H2: Modelagem temporal: Avaliar impacto temporal requer modelar o intervalo temporal entre recomendações como variável endógena (condicionada ao histórico e às recomendações).

H3: Efeitos do tempo: Ajustar a exploração com base na informação temporal pode trazer ganhos de recompensa, diversidade e cobertura.

H4: Limitações: A incorporação do tempo enfrenta restrições de protocolos padronizados e formas de avaliação adequadas, mas oferece potencial para recomendações mais adaptativas e condizentes com dinâmicas reais.

1.4 Objetivo e Escopo

Sumarizando as lacunas, perguntas de pesquisa e hipóteses, o objetivo desta dissertação é investigar o impacto da incorporação de informação temporal no balanceamento entre exploração e aprofundamento em *Contextual Multi-Armed Bandits* (CMAB) para recomendação. São discutidas as limitações da avaliação *offline* (proposta detalhada no Capítulo 5) e adota-se avaliação *online* (Capítulo 6) em simulação com modelagem explícita do tempo de retorno, alinhada a diretrizes *time-aware*. O escopo inclui a implementação de agentes lineares amplamente usados como *benchmarks* e o método proposto *Time-Aware LinBoltzmann* (Capítulo 7).

1.5 Resultados e Contribuições

As principais contribuições deste trabalho são:

- **Diagnóstico**: evidências experimentais das limitações de avaliações *offline*, que tendem a favorecer políticas gulosas frente à exploração adaptativa.
- **Metodologia**: adaptação do simulador KuaiSim para suportar CMAB, múltiplas sessões e a modelagem explícita do tempo de retorno.
- **Algoritmo**: desenvolvimento do *Time-Aware LinBoltzmann*, um algoritmo exploratório que modula dinamicamente a exploração a partir do intervalo temporal.
- **Análise**: estudo dos efeitos do uso de informação temporal em recompensa, diversidade intra-lista e cobertura.
- **Fundamentos**: diretrizes para futuras pesquisas em CMAB sensíveis ao tempo (*time-aware*), tanto em avaliação quanto em modelagem algorítmica.

1.6 Organização do Trabalho

Esta dissertação está estruturada em nove capítulos. O Capítulo 2 apresenta a fundamentação teórica, abordando sistemas de recomendação, fundamentos de CMAB, algoritmos lineares de referência e métricas de avaliação. O Capítulo 3 discute os trabalhos relacionados, com foco em protocolos de avaliação, exploração em modelos lineares, temporalidade, além de ferramentas e bases de dados. O Capítulo 4 delimita o escopo da

pesquisa, contrastando abordagens de avaliação *offline*, *off-policy* e simulação, e introduz a ideia do método *Time-Aware LinBoltzmann*. Capítulo 5 descreve a avaliação *offline*, incluindo protocolos, algoritmos, dados e resultados, enquanto o Capítulo 6 apresenta a avaliação *online* via simulação, detalhando a base KuaiRand, o simulador KuaiSim, suas modificações e métricas. O Capítulo 7 formaliza o *Time-Aware LinBoltzmann*, detalhando sua formulação, configuração experimental e resultados. O Capítulo 8 discute os achados, limitações e perspectivas futuras, e o Capítulo 9 conclui o trabalho, sintetizando respostas às perguntas de pesquisa e destacando contribuições principais.

Capítulo 2

Fundamentação Teórica

A fundamentação teórica apresenta os principais conceitos e métodos que fundamentam esta pesquisa. Parte-se de uma visão geral dos sistemas de recomendação, apresentando sua motivação, evolução e desafios persistentes. Em seguida, abordam-se os fundamentos teóricos de *Contextual Multi-Armed Bandits* (CMAB), explorando seus algoritmos e métricas.

2.1 Sistemas de recomendação

Sistemas de recomendação são uma resposta prática à massiva quantidade de informação em redes sociais, comércio eletrônico, *streaming*, notícias e turismo. Sua função é selecionar e ordenar itens com base em sinais de preferência (explícitos ou implícitos) e no contexto de uso, buscando maximizar utilidade para cada usuário. Isso envolve modelar usuários e itens, lidar com incerteza e operar em ambientes de grande escala. Ao longo dos anos, a área formou um vocabulário comum e um conjunto de tarefas e dados de referência. *Surveys* e livros consolidaram definições, taxonomias e métodos de avaliação, servindo como base para pesquisa e aplicação industrial (RESNICK; VARIAN, 1997; ADOMAVICIUS; TUZHILIN, 2005; RICCI; ROKACH; SHAPIRA, 2015).

Do ponto de vista de algoritmos, duas categorias tornaram-se centrais: métodos baseados em conteúdo e métodos de filtragem colaborativa. Os primeiros usam atributos de itens e perfis de usuários para estimar relevância, enquanto os segundos aprendem padrões de preferência baseados na matriz usuário–item. Trabalhos pioneiros geravam recomendações baseadas em usuário, encontrando gostos semelhantes para filtrar os itens recomendados, e evidenciando assim o valor da colaboração. Em seguida, modelos baseados em itens trouxeram escalabilidade ao focar vizinhanças de itens e pré-cálculos

eficientes (GOLDBERG et al., 1992; SARWAR et al., 2001). A partir daí, surgiram abordagens híbridas que combinam sinais de conteúdo e interação, seguidas por arquiteturas profundas que capturam relações não lineares e sequenciais, ampliando a capacidade de representar preferências dinâmicas em cenários de grande escala (RICCI; ROKACH; SHAPIRA, 2015; ZHANG et al., 2019).

Mesmo com os avanços recentes em sistemas de recomendação, diversos desafios persistem. Um dos principais é a esparsidade de dados, isto é, a presença de grandes quantidades de itens e usuários para os quais há poucas interações registradas. Essa limitação dificulta a aprendizagem de padrões confiáveis e compromete a capacidade de generalização dos modelos. Outro problema recorrente é o chamado *cold-start*, que ocorre quando novos usuários ou itens entram no sistema e ainda não possuem histórico suficiente para que o algoritmo produza boas recomendações. Além disso, requisitos práticos, como escalabilidade (capacidade de lidar com grandes volumes de dados e usuários) e privacidade, também influenciam as escolhas de modelagem e coleta de dados.

No campo da avaliação, métricas tradicionais de desempenho, como a acurácia, não são suficientes para refletir a qualidade da experiência do usuário nem os impactos de longo prazo das recomendações. Por exemplo, se um sistema recomendar repetidamente apenas itens muito populares, poderá obter altas taxas de clique, mas acabará restringindo a diversidade de consumo e reforçando vieses existentes (HERLOCKER et al., 2004). Nesse sentido, aspectos como diversidade (variedade de itens em uma lista de recomendações), novidade (capacidade de expor o usuário a conteúdos ainda não consumidos) e cobertura (proporção do catálogo efetivamente recomendada) tornam-se métricas essenciais (MCNEE; RIEDL; KONSTAN, 2006; CASTELLS; HURLEY; VARGAS, 2022).

Paralelamente, emergiram discussões sobre reprodutibilidade e comparações justas entre algoritmos. Diversos estudos apontam que ganhos reportados em cenários controlados nem sempre se confirmam quando os métodos são avaliados sob protocolos mais rigorosos ou em ambientes distintos (DACREMA; CREMONESI; JANNACH, 2019). Outro aspecto crítico é que os sistemas de recomendação operam em contextos dinâmicos: as preferências dos usuários se transformam ao longo do tempo, e muitas vezes há *delayed feedback*, isto é, o atraso entre a recomendação de um item e a resposta observada do usuário (por exemplo, quando alguém recebe a indicação de um filme, mas só o assiste dias depois). Além disso, a coleta de dados sofre com o chamado viés de posição, fenômeno em que itens apresentados nas primeiras posições da lista tendem a receber mais cliques, independentemente de sua relevância, o que distorce a avaliação do sistema.

Esses desafios evidenciam que a tarefa de recomendar não é estática, mas sim um processo sequencial, no qual cada decisão influencia as interações futuras. Para lidar com essa natureza dinâmica, torna-se necessário adotar modelos capazes de balancear, em tempo de execução, a exploração e o aprofundamento. Nesse contexto, os CMAB surgem como uma alternativa promissora, oferecendo uma estrutura teórica e prática

para enfrentar esses problemas.

2.2 Fundamentos de *bandits* contextuais (CMAB)

Os *bandits* contextuais são um modelo utilizado para apoiar a tomada de decisões em cenários sequenciais. A ideia central é lidar com o dilema entre exploração (recomendar itens novos ou ainda pouco conhecidos, para coletar mais informações) e aprofundamento (recomendar opções que já se mostraram eficazes no passado). O objetivo é maximizar a recompensa acumulada, ou seja, os benefícios obtidos ao longo do tempo, reduzindo o chamado arrependimento (*regret*). O arrependimento mede a diferença entre a recompensa alcançada e aquela que poderia ter sido obtida se o sistema tivesse sempre feito a melhor recomendação possível (SILVA et al., 2022; BOUNEFFOUF; RISH; AGGARWAL, 2020).

O funcionamento dos CMAB pode ser resumido em duas etapas principais: (i) o algoritmo recebe um contexto, que pode incluir informações sobre o usuário (histórico de interações, perfil demográfico), sobre o item (gênero, popularidade, metadados) ou sobre o ambiente (horário de acesso, dispositivo utilizado); e (ii) a partir desse contexto, a política de recomendação decide quais itens sugerir, atribuindo probabilidades às opções e atualizando suas estimativas conforme observa o retorno do usuário. Essa abordagem é particularmente relevante em sistemas de recomendação, pois nesses ambientes o *feedback* é parcial — o sistema observa apenas a reação do usuário ao item que foi efetivamente exibido, e não às alternativas que poderiam ter sido mostradas (TEKIN; TURğAY, 2018; QASSIMI; RAKRAK, 2025). Em cenários dinâmicos, nos quais os interesses dos usuários mudam ao longo do tempo, torna-se ainda mais importante contar com mecanismos de adaptação contínua, ajustando a estratégia de exploração e aprofundamento em tempo de execução (ZENG et al., 2016). Além disso, a escolha de quais atributos compõem o contexto não é trivial: informações irrelevantes podem introduzir ruído, enquanto variáveis com relação causal direta ao comportamento do usuário podem aumentar a capacidade de generalização e a precisão das recomendações (WU; IYER; WANG, 2018).

O uso de CMAB tem se expandido para problemas cada vez mais complexos. Um exemplo é o de múltiplos objetivos, em que o sistema precisa equilibrar não apenas a precisão, mas também outras dimensões de qualidade, como diversidade e novidade. Outro desafio aparece nos cenários de *slates*, quando a decisão não envolve um único item, mas sim um conjunto de recomendações apresentado em determinada ordem, o que intensifica o efeito do viés de posição (QASSIMI; RAKRAK, 2025). Além disso, permanecem problemas clássicos como o *cold-start* (ausência de histórico para novos usuários ou itens), a esparsidade dos dados e a necessidade de considerar múltiplos *stakeholders* com objetivos potencialmente conflitantes, o que torna indispensável o uso de estratégias de balanceamento e de mecanismos de governança de risco (LETARD et al., 2024; LACERDA, 2017).

Outro ponto importante diz respeito aos chamados cenários de alta dimensionalidade, nos quais o contexto disponível contém um número muito grande de variáveis, por exemplo, quando cada item possui descrições textuais extensas, imagens ou centenas de atributos, ou quando os usuários geram sinais complexos de interação em diferentes dispositivos e momentos (BAN; HE; COOK, 2021). Nesses casos, métodos lineares podem ter dificuldade em representar adequadamente todas as combinações possíveis de atributos, sofrendo com a chamada "maldição da dimensionalidade" (SHI et al., 2023). Para lidar com esse desafio, pesquisas recentes exploram abordagens baseadas em redes neurais e métodos adaptativos, capazes de extrair representações mais compactas e capturar relações não lineares entre contexto, item e recompensa (WANG; SHI; LUO, 2025).

Por fim, observa-se também um interesse crescente em selecionar atributos de maneira mais criteriosa, seja com apoio de especialistas de domínio ou critérios de causalidade. Essa seleção cuidadosa pode reduzir ruídos, evitar correlações ruins e melhorar a capacidade de generalização dos modelos (LEE; SIEDAHMED; HEFFERNAN, 2024; GUTOWSKI et al., 2018). As aplicações de CMAB já alcançam diferentes áreas, como comércio eletrônico, saúde, educação e turismo, cada uma trazendo requisitos próprios em termos de risco, latência e interpretabilidade (GAN; KWON, 2022; YU et al., 2024; AMEKO et al., 2020; NGUYEN et al., 2023; Chen, Yizhe, 2025; MCINERNEY et al., 2018).

Esses avanços mostram que os *bandits* contextuais não são apenas um modelo teórico, mas sim uma ferramenta prática que pode ser adaptada a diferentes cenários. No entanto, para que cumpram seu papel de equilibrar exploração e aprofundamento, é necessário escolher uma estratégia de decisão adequada. Entre as várias possibilidades, um grupo de grande destaque são os métodos lineares. Eles partem da suposição de que a recompensa pode ser explicada por uma relação aproximadamente linear entre o contexto do usuário e as características dos itens. Essa família de algoritmos é amplamente utilizada em sistemas de recomendação porque combina boa eficiência computacional com resultados competitivos em termos de recompensa e arrependimento.

A seguir, apresentam-se os modelos lineares (**Lin**, **LinGreedy**, **LinUCB** e **LinTS**) juntamente com a formulação matemática que os fundamenta.

2.2.1 Formulação do Problema

O problema é formulado dentro do cenário de CMAB. Para compreendê-lo, considera-se que um agente precisa tomar uma decisão repetidamente ao longo do tempo.

A cada instante $t = 1, 2, \ldots, T$, ocorre o seguinte:

1. O agente observa um conjunto de opções (também chamadas de braços, do inglês *arms*), representado por $\mathcal{A}$.

2. Para cada braço $a \in \mathcal{A}$, há um vetor de contexto $\mathbf{x}_{t,a} \in \mathbb{R}^d$, que contém informações sobre o estado atual. Esse contexto pode incluir, por exemplo, características do usuário e do item a ser recomendado.

3. O agente deve escolher um braço $a_t \in \mathcal{A}$.

4. Após a escolha, recebe uma recompensa $r_{t,a_t} \in \mathbb{R}$, que representa o retorno observado (em sistemas de recomendação, isso pode ser um clique, uma curtida, ou qualquer sinal de engajamento).

O objetivo final é maximizar a soma das recompensas acumuladas ao longo do tempo:

$$\sum_{t=1}^{T} r_{t,a_t},$$

2.2.2 Modelo Linear

Para tratar o problema, assume-se um modelo linear para a recompensa esperada de cada braço. Em outras palavras, considera-se que, dado um contexto $\mathbf{x}_{t,a}$, a recompensa média que se espera obter segue a seguinte formulação:

$$\mathbb{E}[r_{t,a} \mid \mathbf{x}_{t,a}] = \mathbf{x}_{t,a}^{\top} \boldsymbol{\theta}_a, \tag{1}$$

em que:

- $\mathbf{x}_{t,a} \in \mathbb{R}^d$: vetor de contexto associado ao braço a;
- $\boldsymbol{\theta}_a \in \mathbb{R}^d$: vetor de parâmetros desconhecidos, específico de cada braço.

Esse vetor $\boldsymbol{\theta}_a$ captura como cada dimensão do contexto impacta a recompensa esperada daquele braço.

Ao assumir um modelo linear por braço, evita-se que os parâmetros de diferentes braços interfiram uns nos outros. Isso significa que cada opção pode ter sua própria relação com o contexto, permitindo heterogeneidade na forma como recompensas são geradas. Por exemplo, dois filmes podem responder de forma muito diferente às mesmas características de um usuário, e esse modelo linear separado permite capturar essas diferenças.

2.2.3 Modelo Linear Incremental por Braço

Todos os algoritmos apresentados nesta seção compartilham a mesma base: um modelo linear incremental associado a cada braço. Esse modelo tem como objetivo estimar, a partir dos contextos já observados, os parâmetros que melhor explicam as recompensas recebidas.

2.2.4 Construção das estatísticas

Para cada braço a, são mantidas duas estruturas principais:

- uma matriz $\mathbf{A}_a \in \mathbb{R}^{d\times d}$;
- um vetor $\mathbf{b}_a \in \mathbb{R}^d$.

Essas duas estruturas são atualizadas de forma incremental, isto é, a cada nova interação. Elas correspondem ao resultado da regressão linear com regularização.

A inicialização é feita da seguinte maneira:

$$\mathbf{A}_a \leftarrow \lambda \mathbf{I}_d, \quad \mathbf{b}_a \leftarrow \mathbf{0}_d,$$

onde $\lambda > 0$ é um parâmetro de regularização que garante estabilidade numérica, e $\mathbf{I}_d$ é a matriz identidade de dimensão d.

Atualização a cada interação

Sempre que o braço a_t é selecionado no instante t, e o agente observa o par formado pelo contexto $\mathbf{x}_{t,a_t}$ e pela recompensa recebida r_{t,a_t}, realiza-se a seguinte atualização:

$$\mathbf{A}_{a_t} \leftarrow \mathbf{A}_{a_t} + \mathbf{x}_{t,a_t}\mathbf{x}_{t,a_t}^\top, \tag{2}$$

$$\mathbf{b}_{a_t} \leftarrow \mathbf{b}_{a_t} + r_{t,a_t}\,\mathbf{x}_{t,a_t}. \tag{3}$$

Intuitivamente:

- a matriz $\mathbf{A}_a$ acumula informações sobre os contextos já observados;
- o vetor $\mathbf{b}_a$ acumula as mesmas informações ponderadas pelas recompensas.

Estimativa dos parâmetros

Com as estatísticas atualizadas, a estimativa do vetor de parâmetros do braço a é obtida resolvendo o sistema linear:

$$\widehat{\boldsymbol{\theta}}_a = \mathbf{A}_a^{-1}\mathbf{b}_a. \tag{4}$$

Este vetor $\widehat{\boldsymbol{\theta}}_a$ é a aproximação do verdadeiro parâmetro desconhecido $\boldsymbol{\theta}_a$. Ele captura como cada dimensão do contexto impacta na recompensa média esperada do braço a.

Atualização eficiente da inversa

Na prática, calcular a inversa da matriz $\mathbf{A}_a$ a cada passo é custoso. Por isso, costuma-se manter diretamente a inversa $\mathbf{A}_a^{-1}$ e atualizá-la de forma incremental, utilizando a fórmula de Sherman–Morrison (SHERMAN; MORRISON, 1950):

$$\mathbf{A}_{a_t}^{-1} \leftarrow \mathbf{A}_{a_t}^{-1} - \frac{\mathbf{A}_{a_t}^{-1}\mathbf{x}_{t,a_t}\mathbf{x}_{t,a_t}^{\top}\mathbf{A}_{a_t}^{-1}}{1+\mathbf{x}_{t,a_t}^{\top}\mathbf{A}_{a_t}^{-1}\mathbf{x}_{t,a_t}} \tag{5}$$

Dessa forma, a inversa é ajustada de maneira eficiente a cada nova interação, sem necessidade de recalcular a matriz inteira.

Predição e incerteza

Dada a estimativa dos parâmetros, o valor predito para o par (t, a) é:

$$\widehat{r}_{t,a} = \mathbf{x}_{t,a}^{\top}\widehat{\boldsymbol{\theta}}_a.$$

Além disso, pode-se calcular uma medida de incerteza local, também conhecida como raio de confiança:

$$s_{t,a} = \sqrt{\mathbf{x}_{t,a}^{\top}\mathbf{A}_a^{-1}\mathbf{x}_{t,a}}. \tag{6}$$

Essa medida indica o quanto o modelo ainda está incerto em relação àquele contexto. Braços com contextos pouco explorados tendem a apresentar maior incerteza.

2.2.5 Estratégias de Seleção de Ação

A base linear descrita anteriormente é comum a todos os algoritmos. A diferença entre eles surge apenas no momento da seleção de ações, isto é, na regra utilizada para decidir qual braço será escolhido a cada instante.

Nesta dissertação, são comparadas quatro políticas distintas, que atuam como as políticas base e de referência para posterior comparação com o método temporal:

- ❑ **Lin** (puramente determinístico);
- ❑ **LinGreedy** (ε-greedy);
- ❑ **LinUCB** (Upper Confidence Bound linear);
- ❑ **LinTS** (Thompson Sampling linear).

Cada uma dessas políticas utiliza a estimativa $\widehat{\boldsymbol{\theta}}_a$ (Equação 4) como base, mas adota uma estratégia própria para balancear exploração e aprofundamento.

As próximas subseções detalham cada uma dessas abordagens.

2.2.5.1 Lin (Puramente Guloso)

O primeiro algoritmo, denominado aqui de **Lin**, não corresponde a um método difundido na literatura, mas sim a uma simplificação proposta neste trabalho para fins de comparação. Trata-se da forma mais primitiva de um *bandit* linear: um modelo de regressão linear incremental, utilizado de maneira estritamente gulosa, sem qualquer mecanismo explícito de exploração.

Em outras palavras, o **Lin** consiste apenas em estimar os parâmetros lineares de cada braço a partir dos contextos já observados e, em seguida, selecionar de forma determinística o braço com maior valor predito:

$$a_t \;=\; \arg\max_{a \in \mathcal{A}} \; \mathbf{x}_{t,a}^{\top} \widehat{\boldsymbol{\theta}}_a. \tag{7}$$

Isso significa que, dado o contexto observado, calcula-se para cada braço o valor estimado $\widehat{r}_{t,a}$. O braço com maior valor é sempre selecionado.

Uma característica importante é que esse algoritmo não introduz exploração de forma explícita. Qualquer variação inicial que leve o agente a tentar diferentes braços ocorre apenas devido às incertezas presentes no começo do aprendizado, quando os parâmetros ainda não estão bem ajustados. Conforme o modelo se estabiliza, a tendência é sempre recomendar os mesmos braços considerados mais promissores.

Esse comportamento torna o **Lin** um bom *baseline*, pois mostra o efeito de um agente que se aprofunda rapidamente nas opções mais recompensadoras conhecidas, mas que pode deixar de descobrir alternativas potencialmente melhores.

2.2.5.2 LinGreedy (ε-LinGreedy)

O algoritmo **LinGreedy** (SUTTON; BARTO, 1998), também conhecido como ε-LinGreedy, introduz de forma simples a ideia de exploração controlada.

A lógica é a seguinte:

- com probabilidade $1-\varepsilon$, o agente escolhe o braço guloso, ou seja, aquele com maior valor predito, exatamente como no Lin;
- com probabilidade ε, o agente escolhe aleatoriamente qualquer braço do conjunto $\mathcal{A}$, sem levar em conta o valor predito.

Formalmente:

$$a_t = \begin{cases} \arg\max_{a \in \mathcal{A}} \; \mathbf{x}_{t,a}^{\top} \widehat{\boldsymbol{\theta}}_a, & \text{com prob. } 1-\varepsilon, \\ \text{um braço escolhido de forma uniforme em } \mathcal{A}, & \text{com prob. } \varepsilon. \end{cases} \tag{8}$$

O parâmetro $\varepsilon \in [0,1]$ controla a intensidade da exploração:

- ❏ valores pequenos de ε tornam o comportamento mais próximo ao Lin, focado no aprofundamento;
- ❏ valores maiores levam a mais tentativas aleatórias, aumentando a chance de explorar novas opções.

É importante notar que a exploração em LinGreedy não é guiada pela incerteza, ou seja, a escolha aleatória não leva em conta se um braço já foi pouco testado ou não; simplesmente seleciona de forma uniforme entre todas as opções. Apesar de simples, esse mecanismo garante que, no longo prazo, todos os braços tenham alguma chance de serem avaliados.

2.2.5.3 LinUCB

O algoritmo **LinUCB** (*Linear Upper Confidence Bound*) (LI et al., 2010) introduz um mecanismo de exploração guiada pela incerteza.

A ideia é simples: ao invés de usar apenas o valor predito pelo modelo linear, o algoritmo adiciona a esse valor um bônus de confiança que depende da incerteza local associada ao braço.

Esse valor ajustado é definido como:

$$p_{t,a} \;=\; \mathbf{x}_{t,a}^{\top}\widehat{\boldsymbol{\theta}}_a \;+\; \alpha\, s_{t,a}, \qquad a_t \;=\; \arg\max_{a\in\mathcal{A}}\; p_{t,a}, \tag{9}$$

em que:

- ❏ o primeiro termo $\mathbf{x}_{t,a}^{\top}\widehat{\boldsymbol{\theta}}_a$ corresponde ao valor predito da recompensa;
- ❏ o segundo termo $\alpha\, s_{t,a}$ é o bônus de confiança, onde $s_{t,a}$ representa a incerteza local (Equação 6) e $\alpha \geq 0$ é um hiperparâmetro que regula o grau de exploração.

Intuitivamente, o LinUCB favorece braços que apresentam maior incerteza. Se um braço nunca foi muito testado, seu raio de confiança $s_{t,a}$ tende a ser alto, aumentando sua chance de ser selecionado. Assim, esse método equilibra exploração e aprofundamento de maneira mais sofisticada do que o LinGreedy, pois a exploração deixa de ser aleatória e passa a ser informada pela incerteza.

2.2.5.4 LinTS (Linear Thompson Sampling)

O algoritmo **LinTS**, ou **Linear Thompson Sampling** (AGRAWAL; GOYAL, 2013), utiliza princípios bayesianos para balancear exploração e aprofundamento, por meio da técnica chamada correspondência de probabilidade.

A lógica é a seguinte: em vez de escolher sempre o braço com maior valor predito, o algoritmo considera a incerteza dos parâmetros estimados e, em cada rodada, amostra um conjunto possível de parâmetros para cada braço a partir de uma distribuição probabilística.

Formalmente, para cada braço a, amostra-se:

$$\widetilde{\boldsymbol{\theta}}_a \;\sim\; \mathcal{N}\Big(\widehat{\boldsymbol{\theta}}_a,\; \nu^2 \mathbf{A}_a^{-1}\Big)\,, \tag{10}$$

onde:

- $\widehat{\boldsymbol{\theta}}_a$ é a estimativa atual dos parâmetros;
- $\mathbf{A}_a^{-1}$ representa a incerteza do modelo;
- $\nu^2 > 0$ é um hiperparâmetro que controla a variância da amostragem e, portanto, a intensidade da exploração.

Com esse vetor amostrado $\widetilde{\boldsymbol{\theta}}_a$, calcula-se o valor predito para cada braço:

$$a_t \;=\; \arg\max_{a \in \mathcal{A}} \; \mathbf{x}_{t,a}^{\top} \widetilde{\boldsymbol{\theta}}_a. \tag{11}$$

A escolha do braço é, assim, resultado de uma competição entre valores amostrados, e não apenas dos valores fixos estimados pelo modelo.

Esse mecanismo gera um equilíbrio natural:

- braços com recompensas mais conhecidas têm maior chance de serem escolhidos;
- mas braços com alta incerteza também podem ser selecionados, já que sua distribuição amostral é mais ampla.

Dessa forma, o LinTS explora e aprofunda de maneira estocástica, adaptando-se à incerteza de forma probabilística.

2.2.6 Síntese Comparativa

Após detalhar cada um dos quatro algoritmos, é útil organizar suas características de forma resumida. A Tabela 1 apresenta, em formato de tabela, os mecanismos de exploração utilizados, os hiperparâmetros envolvidos e a regra de decisão de cada método.

Observações

(i) Todos os métodos compartilham a mesma estimação linear incremental apresentada em subseção 2.2.3. A diferença entre eles está apenas na estratégia de seleção de ações.

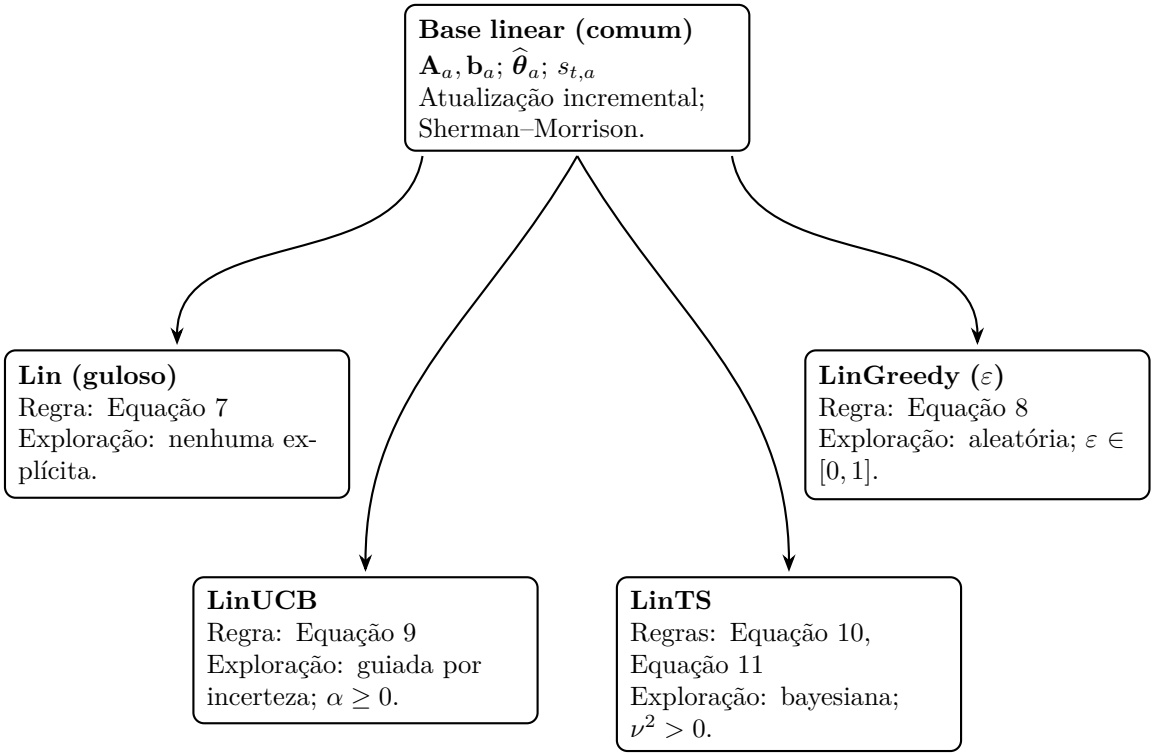
(ii) Os hiperparâmetros ε, α e ν^2 impactam o grau de exploração.

(iii) Em aplicações *online*, é fundamental que o algoritmo consiga operar em tempo real. Para isso, manter a matriz inversa $\mathbf{A}_a^{-1}$ atualizada por meio da fórmula de Sherman–Morrison (Equação 5) reduz significativamente o custo computacional, tornando o aprendizado incremental viável em escala.

Tabela 1 – Resumo comparativo dos principais algoritmos lineares de *Multi-Armed Bandits*, com indicação do tipo de exploração, hiperparâmetros e regra de decisão.

Método	**Exploração**	**Hiperparâmetro(s)**	**Regra de decisão**
Lin	Nenhuma	—	$\arg\max_a \ \mathbf{x}_{t,a}^\top \widehat{\boldsymbol{\theta}}_a$
LinGreedy	Aleatória uniforme	$\varepsilon \in [0,1]$	Guloso $(1-\varepsilon)$ ou uniforme em $\mathcal{A}$ (ε)
LinUCB	Bônus de confiança	$\alpha \geq 0$	$\arg\max_a \ \mathbf{x_{t,a}}^\top \widehat{\boldsymbol{\theta}}_a + \alpha s_{t,a}$
LinTS	Amostragem Bayesiana	$\nu^2 > 0$	$\widetilde{\boldsymbol{\theta}}_a \sim \mathcal{N}(\widehat{\boldsymbol{\theta}}_a,\ \nu^2 \mathbf{A}_a^{-1})$; $\arg\max_a \ \mathbf{x}_{t,a}^\top \widetilde{\boldsymbol{\theta}}_a$

Figura 1 – Mesma base linear para as quatro estratégias.



Fonte: autoria própria.

Em síntese, todos os algoritmos partem da mesma base linear e diferem apenas na forma como equilibram exploração e aprofundamento, como demonstrado na Figura 1. Esse contraste torna esses quatro métodos referências naturais como *benchmarks* para avaliar novas propostas.

2.2.7 Extensões e Avanços dos Bandits Lineares

Extensões recentes dos *bandits* lineares buscam maior expressividade e controle. Modelos como o **DeepLinUCB** combinam redes neurais para aprender representações mais

ricas dos dados com uma camada linear final que mantém a capacidade de quantificar incerteza, unindo flexibilidade e interpretabilidade local (SHI et al., 2023; XU et al., 2022). Esquemas adaptativos (por exemplo, **AdaLinUCB** e variantes **CLUCB2**) ajustam automaticamente o nível de exploração de acordo com a disponibilidade de dados, restrições impostas ou mudanças no ambiente (GUO; WANG; LIU, 2019; GARCELON et al., 2020). Também se destacam técnicas de privacidade e aprendizado federado, que permitem treinar modelos de forma distribuída e com menor risco de vazamento de informações sensíveis (CAO et al., 2023).

2.3 Métricas de avaliação

A avaliação de sistemas de recomendação vai além da acurácia pontual, pois diferentes métricas capturam aspectos complementares da experiência do usuário. Nesta seção, são descritas as principais métricas utilizadas neste trabalho, com suas respectivas formulações matemáticas.

2.3.1 Acurácia de ranqueamento

A métrica **NDCG** (*Normalized Discounted Cumulative Gain*) (SHANI; GUNAWARDANA, 2011; LACERDA, 2015; FREEMAN; RAWSON, 2021) mede a qualidade da ordenação da lista de recomendação, atribuindo maior peso a itens relevantes que aparecem nas primeiras posições.

O cálculo inicia-se pelo ganho acumulado descontado (*Discounted Cumulative Gain*, DCG), definido como:

$$DCG@k = \sum_{i=1}^{k} \frac{2^{rel_i} - 1}{\log_2(i+1)},$$

em que rel_i é a relevância do item recomendado na posição i e k o tamanho da lista considerada.

Para normalizar, utiliza-se o *Ideal DCG* (IDCG), correspondente ao DCG de uma lista perfeitamente ordenada. Assim, o NDCG é definido como:

$$NDCG@k = \frac{DCG@k}{IDCG@k}.$$

2.3.2 Recompensa

A **recompensa média** captura diretamente a utilidade das recomendações do ponto de vista do usuário. Cada interação gera uma recompensa r_i, e a métrica é calculada como:

$$\text{Recompensa Média} = \frac{\sum_{i=1}^{N} r_i}{N},$$

em que N é o número total de interações observadas (LI et al., 2010; QASSIMI; RAKRAK, 2025; Chen, Yizhe, 2025).

Também pode ser analisada a **recompensa acumulada**, obtida pela soma das recompensas ao longo do tempo, refletindo o progresso de aprendizado da política.

2.3.3 Diversidade

A diversidade busca avaliar até que ponto o sistema de recomendação evita redundâncias e promove a exposição a conteúdos variados. Nesta dissertação, foram consideradas duas métricas complementares: **novidade** e **diversidade intra-lista (ILD)** (SHANI; GUNAWARDANA, 2011).

2.3.3.1 Novidade

A **novidade** quantifica o quão incomuns são os itens recomendados, levando em conta sua popularidade no conjunto de dados. Para um item i, a popularidade $pop(i)$ é definida como a frequência relativa de interações em que o item aparece. A novidade é calculada como:

$$\text{Novidade} = \frac{1}{|R|} \sum_{i \in I} \log_2 pop(i),$$

em que I representa o conjunto de itens recomendados ao longo das R interações, e $|R|$ corresponde ao número total de interações consideradas. Valores mais altos de novidade indicam que o sistema está sugerindo itens menos populares, ampliando a exposição do usuário a conteúdos diferentes dos mais consumidos (SHANI; GUNAWARDANA, 2011).

2.3.3.2 Diversidade Intra-Lista (ILD)

A **Diversidade Intra-Lista**, (*Intra-List Diversity*, ILD) mede a variedade de itens dentro de uma mesma lista de recomendação, buscando evitar redundância entre as opções apresentadas ao usuário.

Considerando um conjunto de k itens distintos $\{i_1, i_2, \ldots, i_k\}$ com vetores de representação normalizados v_j, a ILD é definida como:

$$\text{ILD} = 1 - \frac{2}{k(k-1)} \sum_{j=1}^{k} \sum_{l=j+1}^{k} \cos(v_j, v_l),$$

em que $\cos(v_j, v_l)$ representa a similaridade de cosseno entre os vetores dos itens i_j e i_l (MEHROTRA; XUE; LALMAS, 2020; HUANG et al., 2022).

Valores mais altos de ILD indicam listas mais heterogêneas, o que favorece a descoberta de novos conteúdos e pode ampliar a satisfação de usuários com múltiplos interesses.

2.3.4 Cobertura

A **cobertura** mede a proporção do catálogo de itens que foi efetivamente exposta aos usuários.

$$\text{Cobertura} = \frac{|I_{\text{expostos}}|}{|I_{\text{total}}|},$$

em que $|I_{\text{expostos}}|$ corresponde ao número de itens distintos recomendados pelo menos uma vez e $|I_{\text{total}}|$ ao número total de itens do catálogo (SHIMIZU et al., 2024).

Além disso, a métrica de **área sob a curva de cobertura** (CoverageAUC) captura a dinâmica temporal da exploração:

$$\text{CoverageAUC} = \frac{1}{T}\sum_{t=1}^{T}\frac{|I_{\text{expostos}}^{(t)}|}{|I_{\text{total}}|},$$

em que T é o número de passos da simulação e $|I_{\text{expostos}}^{(t)}|$ o número acumulado de itens distintos até o passo t.

Essa métrica diferencia algoritmos que chegam à mesma cobertura final, mas em ritmos distintos.

2.3.5 Engajamento e temporalidade

Além de acurácia, diversidade e cobertura, são empregadas métricas voltadas a engajamento e dinâmica temporal:

2.3.5.1 Tamanho médio da sessão

Mede o número médio de interações por sessão de consumo dos itens, refletindo o nível de engajamento do usuário (ZHAO et al., 2023):

$$\text{Tamanho Médio da Sessão} = \frac{1}{S}\sum_{s=1}^{S} n_s,$$

em que S é o número de sessões e n_s o número de interações na sessão s.

2.3.5.2 Tempo médio de retorno

Avalia a frequência de retorno dos usuários ao sistema, a partir dos intervalos temporais (δ_t) entre sessões consecutivas (ZHAO et al., 2023):

$$\text{Tempo Médio de Retorno} = \frac{1}{N}\sum_{i=1}^{N}\delta_t^{(i)},$$

em que N é o número total de retornos observados.

Essa métrica adiciona uma perspectiva temporal à avaliação, fundamental em cenários em que o intervalo entre consumos influencia o equilíbrio entre exploração e aproveitamento.

2.3.6 Justiça e equilíbrio

Por fim, destacam-se métricas relacionadas à **justiça** (*fairness*), que avaliam se diferentes grupos de usuários ou itens recebem exposição de maneira equilibrada. O índice de Gini generalizado mede desigualdade na distribuição de exposição, enquanto o *Pareto regret* quantifica o equilíbrio entre múltiplos objetivos, como acurácia e diversidade (LACERDA, 2015; JEUNEN; GOETHALS, 2021; HUANG et al., 2021; QASSIMI; RAKRAK, 2025).

2.4 Síntese

Este capítulo apresentou os fundamentos teóricos que sustentam esta pesquisa. Iniciou-se com uma visão geral dos sistemas de recomendação e de seus principais desafios, como esparsidade, *cold-start*, viés de posição, *delayed feedback* e variações temporais nas preferências dos usuários, destacando a importância de equilibrar exploração e aprofundamento em decisões sequenciais.

Em seguida, foram introduzidos os CMAB como modelo adequado para esse equilíbrio, detalhando-se a formulação linear incremental por braço e as estratégias de decisão adotadas pelos algoritmos de referência: **Lin**, **LinGreedy**, **LinUCB** e **LinTS**. Esses métodos compartilham a mesma base linear e diferem apenas na forma como introduzem exploração, servindo como *benchmarks* para a proposta deste trabalho.

Por fim, foram descritas as métricas utilizadas para avaliação: acurácia de ranqueamento (NDCG@k), recompensa média e acumulada, diversidade (novidade e ILD), cobertura (incluindo *CoverageAUC*), engajamento (tamanho médio da sessão), temporalidade (tempo médio de retorno) e justiça.

Essa fundamentação consolida o arcabouço conceitual e metodológico necessário para o desenvolvimento e avaliação de métodos que utilizam informações temporais para ajustar dinamicamente o nível de exploração em sistemas de recomendação.

Capítulo 3

Trabalhos relacionados

Este capítulo contextualiza estudos diretamente relacionados ao escopo deste trabalho. A organização segue três perguntas práticas: (i) como medir o desempenho de políticas de CMAB, (ii) como decidir em face da incerteza usando modelos lineares, (iii) por que e como o tempo altera o problema de recomendação. Depois, são apresentadas ferramentas e bases de dados, finalizando com uma síntese crítica e identificação de lacunas de pesquisa.

3.1 Protocolos de avaliação (como medir)

A literatura usa quatro abordagens de avaliação: *offline*, *online*, simulação e arranjos híbridos (SILVA et al., 2022). Em termos simples, a avaliação *offline* usa registros históricos para estimar como outra política teria se comportado; a *online* mede impacto real com usuários; a simulação cria um ambiente controlado para testar hipóteses; e a abordagem híbrida combina etapas, buscando rapidez e validade externa.

Na OPE, destacam-se três estimadores principais:

- **DM** (*Direct Method*): modela diretamente a recompensa a partir dos dados observados;
- **IPS** (*Inverse Propensity Scoring*): repondera cada interação de acordo com a probabilidade do item ter sido mostrado;
- **DR** (*Doubly Robust*): combina os dois estimadores anteriores, buscando reduzir viés e variância.

Além desses, há também o procedimento de *replay*, que consiste em contabilizar apenas os casos em que a política-alvo teria feito a mesma escolha registrada no *log*. Embora

Tabela 2 – Abordagens de avaliação em CMAB, com vantagens, limites e papel típico.

Abordagem	**Vantagens**	**Limites**	**Papel típico**
Offline (OPE)	Rápida, reprodutível, sem risco para usuários	Sofre com *feedback* parcial e baixa cobertura; viés—variância[a]	Triagem inicial e comparação ampla
Online	Validade externa e efeitos de segunda ordem[b]	Maior custo e risco; precisa de *ramp up*[c] e *guardrails*[d]	Confirmação em produção
Simulação	Controle de cenários e reprodutibilidade[e]	Realismo depende do simulador e das suposições adotadas	Exploração de hipóteses e estresse
Híbrida	Equilíbrio entre rapidez e validade externa[f]	Maior complexidade operacional	*Pipeline* de ponta a ponta

[a] Ver, por exemplo, (LI et al., 2010; KIYOHARA; NOMURA; SAITO, 2024).
[b] (GANGAN; KUDUS; ILYUSHIN, 2021).
[c] Introdução gradual da política em produção, aumentando a exposição conforme sua segurança é confirmada.
[d] Métricas de segurança que atuam como limites mínimos para evitar degradação significativa da experiência do usuário.
[e] (CAÑAMARES; REDONDO; CASTELLS, 2019).
[f] (HAN, 2024).

simples e não enviesado, o *replay* sofre de alta variância e baixa cobertura, sendo considerado mais um ponto de partida histórico do que um estimador avançado (LI et al., 2010; SWAMINATHAN; JOACHIMS, 2015; KIYOHARA; NOMURA; SAITO, 2024).

Em avaliações *online*, usam-se testes A/B e *interleaving*. O primeiro consiste em dividir os usuários em grupos aleatórios, cada um exposto a uma política diferente, possibilitando comparar seu impacto em métricas globais de forma controlada. Já o *interleaving* mistura resultados de duas políticas em uma mesma lista de recomendações apresentada ao usuário, permitindo comparações diretas com menor variância e necessidade de menor amostra (LEQI et al., 2023). A simulação atua como camada intermediária: permite repetir o mesmo cenário e testar eventos raros antes de arriscar em produção (CAÑAMARES; REDONDO; CASTELLS, 2019). *Pipelines* híbridos combinam essas camadas: triagem *offline*, testes em simulador e validação *online* (HAN, 2024).

Quando os *logs* vêm de políticas pouco exploratórias, a OPE tende a subestimar políticas que exploram mais e a favorecer políticas gulosas. Essa observação orienta o uso de simulação com dinâmica temporal para comparações mais justas.

3.2 Exploração em modelos lineares (como decidir)

Modelos lineares assumem que a recompensa esperada pode ser aproximada por uma combinação linear do contexto. Dentro dessa família, há estratégias de exploração com intuição simples. O LinUCB adiciona um "bônus de confiança" em contextos com pouca

evidência, o que o torna mais otimista quando não sabe o suficiente. O LinTS amostra parâmetros plausíveis dado o histórico, o que o faz explorar mais quando há incerteza e consolidar escolhas quando as evidências aumentam. Já ε-greedy e *softmax* (Boltzmann) injetam aleatoriedade diretamente na escolha das ações, de forma uniforme ou proporcional ao valor estimado (BASTANI; BAYATI; KHOSRAVI, 2020).

A literatura recente combina essas ideias com mais expressividade ou com adaptação: arquiteturas que aprendem representações antes da camada linear (por exemplo, DeepLinUCB) e esquemas que ajustam exploração conforme a quantidade e a qualidade dos dados (por exemplo, AdaLinUCB, CLUCB2) (SHI et al., 2023; GUO; WANG; LIU, 2019). Em ambientes dinâmicos, taxas fixas de exploração costumam ser subótimas; é melhor que o nível de exploração responda ao momento, à incerteza e ao custo esperado de errar (HAO; LATTIMORE; SZEPESVÁRI, 2020).

Dentro da família linear, controlar a exploração de forma sensível ao contexto é central. Esse controle pode ocorrer via probabilidades de Boltzmann, que são moduladas por sinais adicionais do ambiente.

3.3 Temporalidade (por que o tempo importa)

O comportamento do usuário muda com o tempo. Interações antigas perdem valor informativo, há ciclos de consumo e há intervalos entre sessões que afetam o engajamento. Um usuário que retorna após 30 minutos costuma ter um ritmo e um apetite por exploração diferentes de quem retorna após 30 dias. Por isso, muitos trabalhos incorporam tempo na modelagem: pesos de recência, funções de decaimento, modelos em tempo contínuo e arquiteturas neurais com *embeddings* temporais; alguns também ajustam o viés de posição conforme o tempo (YE et al., 2020; FAN et al., 2021; ZHANG et al., 2023; HUANG et al., 2023).

Em CMAB, o tempo altera tanto o contexto quanto a incerteza. Propostas específicas introduzem o tempo no cálculo do bônus (ou da variância) e no próprio mecanismo de exploração: variantes do tipo LinUCB-TA, versões dinâmicas de *Thompson Sampling* e detectores de *drift* que reponderam evidências recentes (KHOSRAVI et al., 2025; YAN et al., 2022; YAN et al., 2023; ZENG et al., 2016). Há também trabalhos que distinguem dinâmicas intra-sessão e inter-sessões, mostrando que os intervalos entre retornos influenciam não apenas a relevância, mas também o custo de explorar novas opções (YOU et al., 2019; LI et al., 2025; CAMPOS; DÍEZ; CANTADOR, 2014; BOGINA et al., 2023).

Sinais temporais, em especial o intervalo entre sessões, aparecem como candidatos naturais para calibrar o nível de exploração em CMAB, com impacto direto nas métricas de curto e de longo prazo.

Com esse panorama, a próxima seção apresenta ferramentas e bases de dados usadas na literatura para operacionalizar esses protocolos e cenários, com destaque para coleções

com ordenação temporal e simuladores que permitem estudar inter-sessões e métricas sensíveis ao tempo.

3.4 Ferramentas e bases de dados

A dimensão temporal é central na avaliação de CMAB. Ignorar sinais como tempo de retorno do usuário, intervalo entre interações e lacunas entre sessões (*session gaps*) pode enviesar estimativas e levar a comparações injustas entre políticas (FILIPOVIC et al., 2021; RABIU et al., 2020; BOGINA et al., 2023). A literatura responde a esse desafio com dois pilares: boas práticas de preparo dos dados com respeito à ordem temporal e ambientes de avaliação capazes de representar dinâmica no tempo (BAO; ZHANG, 2021; DU et al., 2015; YOU et al., 2019).

3.4.1 Boas práticas temporais

O tratamento adequado da dimensão temporal é crucial em sistemas de recomendação, especialmente quando se busca avaliar algoritmos em cenários dinâmicos. Protocolos que ignoram a ordem dos eventos podem levar a conclusões enganosas, seja pela superestimação do desempenho de um modelo, seja pela incapacidade de capturar mudanças reais no comportamento dos usuários ao longo do tempo. Nesse sentido, algumas boas práticas têm sido estabelecidas para reduzir vieses e aumentar a validade dos experimentos:

- **Divisão temporal estrita** (*time-based splitting*): separar treino, validação e teste respeitando a linha do tempo, evitando que interações futuras influenciem o aprendizado.
- **Janelas deslizantes**: reavaliar modelos em janelas móveis, de forma a capturar variações graduais e fenômenos de *concept drift*.
- **Prevenção de vazamento** (*information leakage*): garantir que sinais do futuro não contaminem variáveis de entrada, preservando a causalidade entre contexto e ação.
- **Modelagem sequencial**: quando pertinente, adotar arquiteturas que representem explicitamente a dependência temporal, como mecanismos de atenção, modelos em tempo contínuo ou grafos dinâmicos (YE et al., 2020; YANG; YANG, 2025).

3.4.2 Plataformas e simuladores

Ferramentas de simulação permitem controlar variáveis, repetir cenários raros e testar hipóteses antes da produção. Mesmo com plataformas robustas, permanecem desafios: distância entre simulação e produção, realismo limitado de modelos de usuário, dificuldade de medir efeitos de longo prazo e ausência de métricas subjetivas como satisfação e

confiança (KRAUTH et al., 2020). Efeitos contrafactuais e ciclos de realimentação (*feedback loops*) podem introduzir novos vieses, o que pede protocolos responsáveis e auditorias de risco (HUANG et al., 2020). Na Tabela 3 as ferramentas de simulação mais utilizadas na literatura são elencadas, analisando-se seus pontos fortes e limitações.

Tabela 3 – Plataformas de avaliação com foco temporal.

Ferramenta	**Categoria**	**Pontos fortes**	**Limitações**
RecSim[a]	Simulador modular	Interações sequenciais, agentes configuráveis	Realismo depende do desenho do usuário sintético
RecSim NG[b]	Simulador com programação probabilística	Incerteza explícita, multiagente, cenários complexos	Curva de aprendizado, custo de modelagem
KuaiSim[c]	Simulação com foco em sessões	Dinâmica entre sessões e múltiplos tipos de *feedback*	Cobertura de domínios fora do entretenimento
OBP[d]	*Pipeline* de OPE	Padrões para *off-policy*, protocolos comparáveis	Herda limitações de cobertura dos *logs*
Virtual Taobao[e]	Simulação baseada em dados reais	Escala e realismo de *e-commerce*	Dependência de suposições e de dados proprietários

[a] (IE et al., 2019).
[b] (MLADENOV et al., 2021).
[c] (ZHAO et al., 2023).
[d] (SAITO et al., 2021).
[e] (SHI et al., 2019).

3.4.3 Bases de dados

Há dois tipos principais de conjuntos de dados usados na literatura para CMAB com tempo:

1. **Registros com propensidades** (*bandit logs*): expõem apenas o rótulo da ação escolhida e, quando disponíveis, as probabilidades de exposição (*propensity scores*), permitindo estimadores *off-policy* como IPS e DR (BIETTI; AGARWAL; LANGFORD, 2021; LI et al., 2011; XIE; TANG; ZHU, 2023).

2. **Logs temporais de recomendação**: coleções reais com carimbo de tempo, posição, sessão e metadados de exposição (por exemplo, Yahoo! Front Page), úteis para *replay* e avaliação imparcial de risco (LI et al., 2011).

Coleções clássicas, como MovieLens e Netflix, são frequentemente temporalizadas ou transformadas em ambientes semi-sintéticos para estudar não estacionaridade e efeitos de política (DEREVENTSOV; BIBIN, 2022; RAO, 2020). Para CMAB, certos **metadados**

são essenciais: carimbo de tempo, identificador de sessão, posição do item na lista, identificador de exposição, taxa de exibição por posição e, idealmente, propensidades. Esses campos sustentam OPE coerente e simuladores que respeitam a cronologia (LI et al., 2010).

Observa-se o surgimento de registros mais dinâmicos (Yahoo!, Taobao, Yelp), uma maior atenção à privacidade, e novos *benchmarks* semi-sintéticos para ambientes não estacionários ganham espaço (HE et al., 2020; GU et al., 2024; ZHU et al., 2023). Avançam também métodos de OPE sensíveis ao contexto e a mudanças de política (YAN et al., 2023; ZHENG et al., 2023). Em domínios como saúde, a estrutura temporal é rica e o custo de erro é alto, o que torna esses dados ideais para estudar decisões personalizadas com restrições éticas (VARATHARAJAH; BERRY, 2022).

Apesar dos avanços, ainda faltam padrões amplamente aceitos para divisão temporal, documentação de propensidades e reprodutibilidade (AGARWAL; MCMAHAN; XU, 2023; BETELLO et al., 2025). Isso reforça a importância de *frameworks* abertos e sensíveis ao tempo, que sirvam de base para estudos mais justos e comparáveis.

3.5 Síntese crítica e lacunas de pesquisa

A avaliação *offline* em CMAB normalmente utiliza estimadores *off-policy*, que buscam estimar o desempenho de uma nova política a partir de dados coletados por outra. Esses métodos assumem a condição de *full support*, isto é, que em cada contexto todas as ações tiveram alguma chance de serem escolhidas pela política que gerou os dados.

Quando essa condição não é atendida, surge a chamada *support deficiency*. Esse problema é comum em políticas determinísticas ou muito gulosas, que quase sempre escolhem a mesma ação, especialmente em modelos lineares (LANGFORD; STREHL; WORTMAN, 2008).

Nesses cenários, os estimadores tornam-se instáveis. O método IPS, que repondera cada observação pela probabilidade da ação, sofre com alta variância quando essas probabilidades são muito baixas. O método DR, que combina um modelo de recompensa com ajustes de propensidade, também herda vieses e se torna pouco confiável (WANG; AGARWAL; DUDÍK, 2017; MARY; PREUX; NICOL, 2014).

Mesmo com normalizações, o desalinhamento entre a política-alvo e a política de coleta tende a introduzir viés residual (SAITO; JOACHIMS, 2022; TRAN-THE et al., 2021). Experimentos em dados sintéticos mostram que esse cenário costuma punir políticas exploratórias, que testam novas ações, e dar vantagem aparente a estratégias gulosas (SACHDEVA; SU; JOACHIMS, 2020; STAVINOVA et al., 2022). Isso reforça a necessidade de complementar a avaliação *offline* com simulações.

Outro desafio é o deslocamento de distribuição (*distribution shift*) entre o momento da coleta dos dados e o da implantação da política, ou seja, quando o ambiente ou o

comportamento dos usuários muda ao longo do tempo. Além disso, os *logs* observacionais sofrem com fatores que comprometem a validade causal das estimativas: a influência de variáveis ocultas que afetam simultaneamente o contexto e a recompensa, gerando associações enganosas; o viés de seleção, que decorre do fato de que os dados refletem apenas as ações escolhidas pela política anterior, e não todas as alternativas possíveis; e a presença de dados faltantes, que limita a observação completa do processo. Esses aspectos dificultam a obtenção de conclusões robustas (MU et al., 2022; CHEN et al., 2023; HUANG; WU, 2024).

Para lidar com essas limitações, pesquisas recentes têm investigado diferentes estratégias. Uma linha de trabalho busca decompor o problema em partes menores e mais controláveis, por meio de decomposições fatoradas, de modo a reduzir a complexidade do processo de avaliação. Outras abordagens procuram aumentar a robustez frente a mudanças de distribuição, garantindo que o modelo mantenha bom desempenho mesmo quando o ambiente se altera. Há ainda o uso de aprendizado por imitação, em que a política é treinada para reproduzir decisões observadas em dados históricos, e variantes do estimador DR que incluem ajustes explícitos para reduzir viés e variância (MA; WANG; NARAYANASWAMY, 2019; SAITO; REN; JOACHIMS, 2023). Além disso, técnicas baseadas em *bootstrapping* têm sido aplicadas para obter intervalos de confiança mais confiáveis, enquanto combinações de avaliação *online* e *offline* buscam unir as vantagens de cada paradigma, embora continuem dependendo da qualidade da política de coleta e da estabilidade temporal do ambiente (TRAN-THE et al., 2021; TUCKER; JOACHIMS, 2023).

Nos algoritmos sensíveis ao tempo, surgem inconsistências de protocolo. Há variação em como os dados são divididos (aleatoriamente ou respeitando ordem temporal), nas janelas de avaliação e nas métricas utilizadas, muitas vezes sem considerar frescor ou novidade. Essas diferenças tornam difícil a comparação justa entre métodos (CAMPOS; DÍEZ; CANTADOR, 2014; MEZNI; FAYALA, 2018; SáNCHEZ; BELLOGíN, 2020).

Pequenas mudanças no corte temporal podem até inverter o *ranking* de algoritmos (SCHEIDT; BEEL, 2021). Além disso, problemas como *data leakage* temporal (quando sinais futuros entram no treino) e *model aging* (perda de desempenho por desatualização) raramente são medidos de forma sistemática, comprometendo a reprodutibilidade (JIANG; ZENG, 2022).

Incorporar a dimensão temporal em CMAB traz tanto desafios quanto oportunidades. Entre os principais desafios, destacam-se:

- ❑ adaptar-se a mudanças abruptas ou graduais nas preferências dos usuários (ZENG et al., 2016; WU; LI; WANG, 2019);
- ❑ lidar com *feedback* atrasado, ou seja, situações em que a resposta do usuário ocorre muito tempo depois da recomendação, exigindo técnicas para identificar quais ações

passadas foram responsáveis pelo resultado observado (ZHOU; XU; BLANCHET, 2019);

- ❏ enfrentar esparsidade e *cold-start* em cenários de alta rotatividade (GAN; KWON, 2022; WANG; WANG; HE, 2017; SILVA et al., 2022);

- ❏ lidar com informações temporais incompletas ou parciais sobre o estado do usuário, o que dificulta entender plenamente seu comportamento ao longo do tempo (ZENG et al., 2025; PARK; FARADONBEH, 2022);

Do lado das oportunidades, uma direção promissora é a exploração guiada por sinais temporais, como o intervalo entre sessões de um mesmo usuário, que permite adaptar o nível de exploração conforme o padrão de retorno (ZHANG et al., 2022; PENG et al., 2025). Outra frente de pesquisa envolve algoritmos híbridos que combinam modelos lineares e neurais, aproveitando a interpretabilidade e eficiência dos primeiros com a flexibilidade dos segundos (KHOSRAVI et al., 2025). Também têm sido propostos mecanismos de *hypernetworks*, nos quais uma rede gera os parâmetros de outra, permitindo capturar de forma mais eficaz dependências de curto e longo prazo (SHEN et al., 2023). Além disso, há variações de *Thompson Sampling* que incorporam decaimento temporal, ajustando a incerteza do modelo conforme os dados se tornam mais antigos (YAN et al., 2023). Estratégias mais gerais, como o uso de *ensembles* (combinação de múltiplos modelos), *meta-bandits* (bandits que aprendem a selecionar entre diferentes algoritmos) e modelos autorregressivos (que preveem recompensas a partir de séries temporais), têm mostrado ganhos em estabilidade e na redução do *regret* em cenários não estacionários (QASSIMI; RAKRAK, 2025).

Diante desse panorama, três direções se destacam:

- ❏ Reduzir vieses da avaliação *offline* (*off-policy evaluation*) com protocolos padronizados e relatos completos;

- ❏ Desenvolver *benchmarks* temporais com divisão cronológica e métricas que capturem recência, envelhecimento e estabilidade;

- ❏ Avaliar CMAB sensíveis ao tempo de forma multiobjetivo, indo além da recompensa imediata para incluir diversidade, cobertura e engajamento (BOGINA et al., 2023; HAN, 2024; CAMPOS; DÍEZ; CANTADOR, 2014).

Para garantir avanços consistentes na área, é importante adotar práticas de reprodutibilidade, utilizar *baselines* estáveis e compartilhar a infraestrutura experimental. Protocolos híbridos que integrem avaliação *online* e *offline* podem acelerar experimentos sem comprometer a validade dos resultados (HEJAZINIA et al., 2019; BETELLO et al., 2025). Ainda assim, continua em aberto a questão de como incorporar sinais temporais de forma

sistemática tanto no processo de avaliação quanto no próprio desenho de algoritmos. Essa lacuna abre espaço para novas investigações que explorem estratégias sensíveis ao tempo, capazes de equilibrar exploração e aproveitamento a partir do padrão de retorno dos usuários, direção em que se insere a presente dissertação.

Capítulo 4

Escopo da Pesquisa

Este capítulo apresenta uma visão geral do escopo e abordagem utilizada para investigar o uso de informações temporais como moduladoras da exploração em bandits contextuais aplicados a sistemas de recomendação. O objetivo é expor os elementos centrais da proposta, da motivação ao desenho experimental, e indicar como cada componente se relaciona às perguntas de pesquisa, enquanto os detalhes metodológicos, configurações específicas e resultados quantitativos são tratados nos capítulos seguintes.

A presente pesquisa parte da hipótese de que a informação temporal associada ao comportamento dos usuários pode ser utilizada para modular o nível de exploração em agentes de CMAB. A suposição é que, ao considerar aspectos temporais das interações, o agente pode ajustar dinamicamente sua propensão a explorar novas opções ou a aprofundar preferências já observadas, de modo a equilibrar de forma mais eficaz a relação entre exploração e aprofundamento.

Para investigar essa hipótese, o trabalho foi estruturado em três etapas complementares:

1. **Avaliação offline e off-policy (OPE)**: estabelecendo uma linha de base de desempenho e evidenciando limitações dos protocolos tradicionais de avaliação de CMAB.

2. **Simulação Online**: desenvolvimento e adaptação de um simulador capaz de modelar cenários de recomendação com múltiplas sessões e dinâmicas temporais, permitindo analisar o impacto da informação temporal de forma controlada.

3. **Proposição do método temporal**: implementação de uma variação dos métodos tradicionais de *bandits* lineares, em que a temperatura do processo de exploração é ajustada por *Boltzmann* de acordo com informações temporais derivadas do comportamento dos usuários.

Essa organização em etapas possibilitou, por um lado, compreender e explicitar as limitações das avaliações *offline* e *off-policy* e, por outro, investigar o potencial da informação temporal como elemento regulador da exploração em sistemas de recomendação baseados em CMAB.

Com base nessas três etapas, esta pesquisa busca responder de forma articulada às questões que motivaram sua formulação. As perguntas de pesquisa são retomadas a seguir:

PP1: Como avaliar de maneira justa CMAB com exploração adaptativa? Quais condições e protocolos são mais adequados.

PP2: Como modelar o tempo em CMAB de forma realista? Em especial, como representar o **tempo de retorno** quando ele também é influenciado pelas próprias recomendações.

PP3: Qual o efeito do tempo como fator de exploração? Impactos de usar informações temporais na exploração sobre, por exemplo, recompensa, diversidade e cobertura.

PP4: Quais limitações e desafios práticos dessa abordagem? Restrições de dados, protocolos e generalização.

4.1 Avaliação Offline e Off-Policy

Avaliações *offline* (isto é, usando apenas dados históricos já registrados, sem envolver usuários em tempo real) são de baixo custo, reprodutíveis e adequadas para verificações iniciais de consistência. Nessa fase, examinam-se: (i) as representações de usuários e itens; (ii) os estimadores lineares (modelos simples que combinam características para prever resultados); e (iii) o protocolo de treino e teste (como os dados são divididos).

4.1.1 Viés observado

Em *bandits* contextuais com modelos lineares, esse tipo de avaliação mostrou um padrão sistemático: políticas que "aproveitam o que já funciona" (aprofundam) tendem a parecer melhores do que políticas que "testam alternativas novas" (exploram). Na prática, políticas lineares determinísticas superam consistentemente variantes que incluem exploração explícita. O motivo é que a avaliação *offline* reflete o comportamento da política que gerou os dados (o *logger*, isto é, a política utilizada), assim, ações pouco registradas acabam recebendo pouca ou nenhuma oportunidade na avaliação, o que penaliza métodos que precisam justamente explorar para aprender.

4.1.2 Extensão via OPE

Para ir além desse limite, utiliza-se a avaliação *off-policy* (OPE): estimar, a partir de *logs* existentes, como uma nova política se sairia sem precisar executá-la de fato. Empregou-se o *Open Bandit Pipeline* (OBP), que implementa técnicas amplamente usadas, como IPS, DM e DR. A ideia central é reponderar cada registro do *log* pela "propensão" que a política registradora tinha de escolher aquela ação no contexto observado. Assim, corrige-se parte do viés de se olhar apenas para o que o *logger* fez com mais frequência.

4.1.3 Limitações persistentes

Mesmo com OPE, algumas limitações se mostraram relevantes no cenário estudado:

- **Suporte limitado**: quando o *logger* raramente escolhe certas ações, as propensões associadas a elas tendem a valores próximos de zero, tornando as estimativas instáveis ou pouco informativas.
- **Ignorância da dinâmica temporal**: os estimadores *off-policy* convencionais tratam as interações como independentes, sem considerar que recomendações atuais podem afetar o comportamento futuro do usuário — por exemplo, a probabilidade e o intervalo temporal entre consumos.
- **Desalinhamento de objetivos**: enquanto políticas conscientes de tempo foram concebidas para explorar explicitamente informações temporais e ajustar o nível de exploração ao longo do tempo, a OPE baseia-se apenas na reponderação de *logs* fixos. Dessa forma, não captura os efeitos de realimentação que tais políticas pretendem induzir sobre engajamento e retenção.

4.1.4 Evidência e validação externa

A análise combinada das avaliações *offline* e OPE fundamentou um estudo aceito no *RecSys 2025*, o que reforça a robustez dos achados. Esse resultado evidencia a limitação estrutural de protocolos baseados apenas em dados históricos e motiva a adoção de simulações online na sequência da pesquisa, onde é possível representar de forma explícita a interação entre recomendações, comportamento do usuário e informações temporais.

4.2 Avaliação Online via Simulação

Necessidade da simulação

Para avaliar uma política que modula a exploração em função do tempo e, ao mesmo tempo, altera o comportamento temporal dos usuários, é preciso um ambiente onde re-

comendações e informações temporais (como intervalo entre sessões de consumo) estejam relacionadas. A simulação online permite:

- **Endogeneidade temporal:** informação temporal afeta a política e é afetada por suas decisões.
- **Fechamento do ciclo de feedback:** ações alteram o tempo (intervalo entre sessões), que altera contexto e recompensas futuras.
- **Controle e reprodutibilidade:** parametrização de ruídos, distribuições de retorno e horizontes; *seeds* (sementes definidas para garantir reprodutibilidade) e protocolos padronizados.

A simulação é o primeiro cenário em que a proposta de uso do tempo pode ser julgada pela sua capacidade de aprender via exploração e intervir na dinâmica temporal, algo que nem *offline*, nem OPE capturam adequadamente.

Adaptação do simulador KuaiSim

O ponto de partida foi o *KuaiSim*, um simulador de recomendação de vídeos curtos desenvolvido a partir do conjunto de dados *KuaiRand* (ZHAO et al., 2023). Esse ambiente busca reproduzir de forma controlada a experiência de plataformas de entretenimento que oferecem vídeos em sequência, semelhantes a redes sociais de vídeos curtos. O simulador fornece perfis de usuários, itens de conteúdo, sinais de recompensa (como cliques ou tempo assistido) e também modela o retorno dos usuários ao sistema em diferentes momentos.

Modificações realizadas

Para que o simulador pudesse ser usado no estudo, foi necessário adaptá-lo em alguns aspectos principais:

1. **Agentes de recomendação:** substituiu-se o foco em algoritmos de aprendizado por reforço por *bandits contextuais* lineares. Esses agentes usam informações do contexto (como características do usuário e do item) para decidir o que recomendar. Foi criada uma interface padronizada para treinar, avaliar e escolher recomendações.
2. **Múltiplas sessões:** ampliou-se o suporte para simulações em que o mesmo usuário retorna ao sistema em diferentes sessões. Nesse modelo, as recomendações feitas em um momento podem influenciar quando e como o usuário voltará a interagir, criando um ciclo de dependência entre recomendação e comportamento de retorno (informação temporal).

3. **Protocolos experimentais:** estruturaram-se protocolos que permitem comparar de forma justa diferentes algoritmos de referência (*benchmarks*), como Lin, LinGreedy, LinUCB e LinTS.

Essas adaptações transformaram o *KuaiSim* em um ambiente de experimentação adequado para investigar, em condições controladas e reprodutíveis, o impacto de algoritmos que levam em conta informações temporais no processo de recomendação. Esse passo foi essencial para possibilitar a avaliação da proposta deste trabalho.

4.3 Método Time-Aware LinBoltzmann

O *Time-Aware LinBoltzmann* (também denominado TA-LinBoltzmann ou TALB) é um método proposto de recomendação que combina três ideias simples: (i) um **modelo linear** para estimar, a partir de características do usuário e do item (o *contexto*), a recompensa esperada de cada opção; (ii) uma regra de escolha probabilística chamada seleção por Boltzmann (*softmax*), que transforma pontuações em probabilidades de recomendação; e (iii) um parâmetro de temperatura adaptativa que controla o quanto a política é mais "aleatória" (explora) ou mais "conservadora" (aproveita o que já se sabe), ajustada por informações temporais.

Informações temporais usadas

Denota-se por δ_t o intervalo de tempo entre duas sessões consecutivas do mesmo usuário (isto é, o tempo decorrido desde a última vez que esse usuário interagiu na última sessão até o início da sessão atual). A hipótese investigada é que esse intervalo carrega sinal útil para decidir o nível adequado de exploração.

Componentes do método

1. **Modelo linear**: Cada item a recebe uma recompensa prevista $\hat{r}_a$ com base no contexto (características do usuário e do item). Em termos simples, trata-se de uma combinação linear dessas características para estimar a chance de engajamento (por exemplo, clique ou tempo assistido).

2. **Seleção por Boltzmann (*softmax*)**: Em vez de sempre escolher o item de maior recompensa esperada, gera-se probabilidades de recomendação. Assim, itens com chances menores ainda podem ser escolhidos, permitindo explorar alternativas.

3. **Temperatura adaptativa guiada pelo tempo**: O parâmetro de temperatura $\tau(\delta_t)$ regula o grau de aleatoriedade da escolha: τ alta torna a política mais exploratória; τ baixa a torna mais focada na maior probabilidade. Aqui, τ é ajustada automaticamente como função de δ_t.

Regra de decisão

A probabilidade de recomendar a ação a, dado o contexto x e o intervalo δ_t, é dada por:

$$\pi(a \mid x, \delta_t) = \frac{\exp(\hat{r}_a/\tau(\delta_t))}{\sum_b \exp(\hat{r}_b/\tau(\delta_t))}.$$

Onde $\hat{r}_a$ é a recompensa prevista pelo modelo linear; $\tau(\delta_t)$ é a temperatura que depende do tempo entre sessões; e o denominador garante que as probabilidades somem 1 entre todos os itens candidatos.

Comparação com métodos de referência

Para isolar o efeito da exploração guiada por tempo, o TA-LinBoltzmann foi comparado com métodos lineares conhecidos: **Lin** (estratégia gulosa que sempre aprofunda), **LinGreedy** (*ε-greedy*, que escolhe o melhor na maior parte das vezes e, ocasionalmente, explora ao acaso), **LinUCB** (otimismo nos limites superiores, *Upper Confidence Bound*), e **LinTS** (amostragem de Thompson, *Thompson Sampling*). Todos usam a mesma base linear, diferindo apenas na forma de explorar.

Objetivo e papel no todo

O objetivo não é afirmar superioridade universal, e sim demonstrar como exploração condicionada por informações temporais pode trazer ganhos em cenários nos quais o tempo influencia o comportamento de retorno do usuário. O TA-LinBoltzmann materializa a hipótese central desta dissertação: se o intervalo entre sessões contém informação relevante, então o nível de exploração deve depender do tempo. A avaliação adequada dessa ideia requer ambientes que representem explicitamente a interação entre decisão de recomendação e comportamento futuro do usuário.

4.4 Síntese

Este capítulo delineou o percurso da pesquisa: partiu-se da avaliação *offline*, que revelou vieses estruturais nos protocolos tradicionais, avançou-se para a *off-policy evaluation* (OPE), cuja aplicação mostrou limitações importantes, e culminou-se na adoção de simulações online com o KuaiSim adaptado. Esse ambiente permitiu testar a hipótese central do trabalho por meio do **Time-Aware LinBoltzmann**, método que ajusta a exploração de forma dependente do tempo entre sessões de consumo.

A partir do próximo capítulo, passam a ser detalhadas as implementações realizadas e os resultados obtidos, permitindo avaliar na prática as contribuições propostas.

Capítulo 5

Avaliação Offline

A avaliação é um componente central no desenvolvimento de sistemas de recomendação, pois permite comparar diferentes abordagens de maneira sistemática e reprodutível. No caso de algoritmos de *Contextual Multi-Armed Bandits* (CMAB), esse processo é particularmente desafiador devido à natureza sequencial das interações e à necessidade de equilibrar exploração e aprofundamento.

Este capítulo discute a avaliação *offline* de algoritmos lineares de CMAB, destacando tanto suas vantagens quanto suas limitações. São descritos os algoritmos considerados, os conjuntos de dados empregados, o protocolo experimental adotado e as métricas utilizadas. Em seguida, apresentam-se os resultados obtidos, incluindo uma análise complementar com técnicas de *Off-Policy Evaluation* (OPE). Por fim, evidencia-se como esse tipo de avaliação favorece modelos puramente gulosos, estabelecendo a motivação para a adoção de simuladores interativos, explorados nos capítulos seguintes.

5.1 Avaliação Offline

A forma mais direta de mensurar o desempenho de um agente em sistemas de recomendação seria por meio de experimentos *online*, como testes A/B. Esses protocolos permitem comparar políticas distintas em interação real com usuários, mas envolvem custos elevados, riscos à experiência do usuário e desafios de implementação em larga escala (GILOTTE et al., 2018; AKKER et al., 2024).

Como alternativa, a avaliação *offline* consolidou-se como prática predominante (LI et al., 2010; SILVA et al., 2022). Baseada em registros históricos de interação, essa abordagem — também chamada de *replay evaluation* — simula a execução de um agente sobre dados previamente coletados, respeitando a ordem temporal. Cada recomendação

hipotética é comparada à resposta registrada nos *logs*, possibilitando estimar métricas de qualidade sem interação com usuários reais.

Apesar de sua popularidade, a avaliação *offline* apresenta limitações estruturais. Como os dados estão fixados, o comportamento dos usuários não é influenciado pelas ações do agente, o que compromete a análise de políticas dependentes de exploração. Em consequência, estratégias gulosas tendem a ser artificialmente beneficiadas, enquanto métodos que exploram incertezas são desfavorecidos (DUDIK; LANGFORD; LI, 2011; LI et al., 2011; SAITO et al., 2021; GUPTA et al., 2024).

Dessa forma, esta seção busca oferecer uma visão crítica da avaliação *offline*, preparando o terreno para os resultados apresentados a seguir e para a discussão sobre abordagens de simulação *online*.

5.2 Algoritmos avaliados

A avaliação offline apresentada neste capítulo tem como objetivo comparar diferentes variantes de métodos lineares. Todos os métodos compartilham a mesma estrutura de regressão linear para estimar a recompensa esperada, diferenciando-se apenas na estratégia de seleção de ações, conforme apresentado na seção 2.2.

Foram considerados os seguintes algoritmos:

- **Lin**: estratégia puramente gulosa, equivalente ao método *epsilon-greedy* com $\epsilon = 0$, que seleciona sempre o item de maior recompensa estimada.

- **LinGreedy**: variação do *epsilon-greedy* (SUTTON; BARTO, 1998), em que o parâmetro ϵ controla a frequência de exploração aleatória.

- **LinUCB**: algoritmo baseado no princípio de *Upper Confidence Bound* (UCB) (LI et al., 2010), que adiciona um termo de incerteza à estimativa da recompensa para incentivar a exploração de itens menos conhecidos.

- **LinTS**: versão linear do *Thompson Sampling* (AGRAWAL; GOYAL, 2013), que realiza amostragem bayesiana dos parâmetros do modelo, equilibrando exploração e exploração (*exploration–exploitation trade-off*) de forma probabilística.

Esses algoritmos foram selecionados por representarem os principais métodos lineares utilizados na literatura de CMABs em sistemas de recomendação, permitindo uma análise comparativa das diferentes abordagens de exploração no contexto de avaliação offline (SILVA et al., 2022).

5.3 Conjuntos de dados

Os experimentos de avaliação *offline* foram conduzidos em um conjunto diverso de bases públicas, amplamente utilizadas na literatura de sistemas de recomendação e *bandits* contextuais. A escolha buscou contemplar diferentes domínios (varejo, marcadores sociais, filmes e e-commerce), bem como escalas variadas de usuários/itens e densidades distintas de interação, de modo a avaliar a robustez dos algoritmos em cenários heterogêneos (SILVA et al., 2022).

Panorama da literatura

A Tabela 4 apresenta um levantamento de bases de dados comumente empregadas em estudos de avaliação *offline*. Além do nome da base, indica-se se é de domínio público, o domínio de aplicação e exemplos de trabalhos que a utilizaram.

Esse panorama evidencia a heterogeneidade de domínios e tamanhos das bases utilizadas na literatura. Mais importante, mostra que não há consenso quanto aos protocolos de seleção e pré-processamento: diferentes trabalhos utilizam subconjuntos distintos de usuários e itens, aplicam filtros variados de densidade ou recorte temporal, e adotam métricas divergentes para caracterizar os dados. Essa variabilidade metodológica dificulta comparações diretas entre estudos e pode introduzir vieses que afetam especialmente a avaliação de estratégias de exploração (SILVA et al., 2022).

Bases utilizadas nos experimentos

Para os experimentos desta dissertação, foi utilizado um subconjunto dessas bases, escolhidas de acordo com sua relevância na literatura e disponibilidade pública. Durante o pré-processamento, foram removidas interações duplicadas, eliminados registros inconsistentes e, no caso do *Delicious*, criada uma versão simplificada (*Delicious-PU*), contendo apenas a URL principal de cada item. A Tabela 5 resume as estatísticas dessas bases.

Essas bases oferecem um equilíbrio entre cenários densos (como *MovieLens-100K*) e de grande escala com alta esparsidade (como *Amazon Games* e *RetailRocket*), possibilitando observar como os diferentes algoritmos lineares se comportam em contextos contrastantes.

5.4 Protocolo experimental

A avaliação *offline* foi conduzida seguindo um protocolo unificado, cujo objetivo é garantir reprodutibilidade e comparabilidade entre algoritmos. Esse protocolo, ilustrado na Figura 2, foi inspirado em práticas comuns na literatura de avaliação de bandits contextuais (LI et al., 2010; SILVA et al., 2022).

Tabela 4 – Conjuntos de dados comumente utilizados na avaliação *offline* de *multi-armed bandits*.

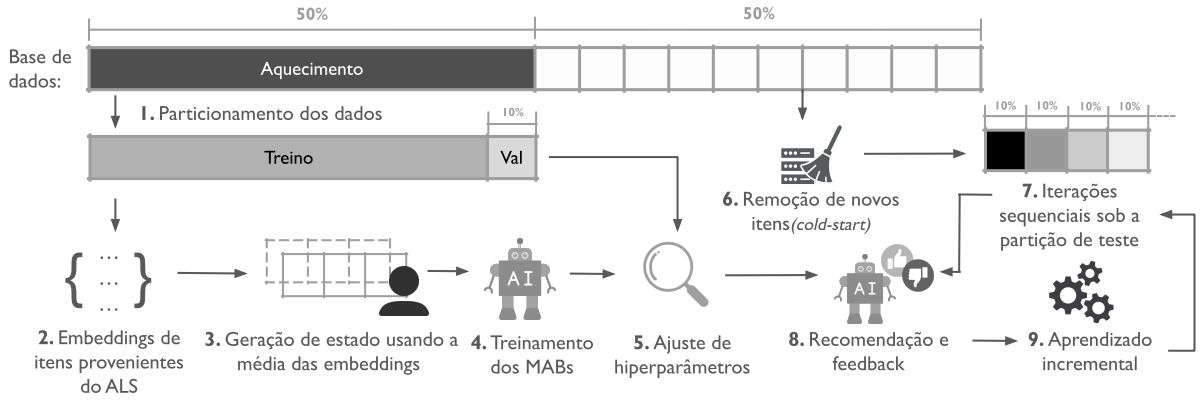
Base de dados	**Público**	**Domínio**
Amazon Review[1]	✓	Varejo
Book-Crossing[2]	✓	Livros
Cheetah Mobile[3]		Artigos
Delicious[4]	✓	Marcadores
GoodReads[5]	✓	Livros
Jester[6]	✓	Piadas
Last.FM[7]	✓	Música
Million Songs[8]	✓	Música
MovieLens[9]	✓	Filmes
NetflixPrize[10]	✓	Filmes
PoliticalNews[11]		Notícias
Spotify[12]		Música
Toutiao[13]	✓	Notícias
Xiami Music[14]		Música
Yahoo! Music[15]	✓	Música
Yahoo! News[16]	✓	Notícias
Yelp[17]	✓	Restaurantes
YOW[18]	✓	Notícias

[1] Chen et al. (2020), Ghoorchian, Kortukov e Maghsudi (2024).
[2] Zhou et al. (2020b).
[3] Liu et al. (2018).
[4] Cesa-Bianchi, Gentile e Zappella (2013), Liu et al. (2018), Jagerman, Markov e Rijke (2019), Nguyen e Lauw (2014), Wang, Wu e Wang (2016).
[5] Chen et al. (2020).
[6] Ghoorchian, Kortukov e Maghsudi (2024).
[7] Cesa-Bianchi, Gentile e Zappella (2013), Caron e Bhagat (2013), Jagerman, Markov e Rijke (2019), Nguyen e Lauw (2014), Wang, Wu e Wang (2016), Wang, Wu e Wang (2017), Xu et al. (2020).
[8] Takemori et al. (2020).
[9] Celis et al. (2019), Chen et al. (2020), Ghoorchian, Kortukov e Maghsudi (2024), Rao (2020), Takemori et al. (2020), Zhou et al. (2020b).
[10] Chen et al. (2020).
[11] Celis et al. (2019).
[12] McInerney et al. (2018).
[13] Zhang et al. (2020).
[14] Zhou et al. (2020a).
[15] Hariri, Mobasher e Burke (2015).
[16] Chapelle e Li (2011), Song, Fragouli e Shah (2019), Tracà, Rudin e Yan (2019), Wang, Wu e Wang (2017), Wu et al. (2017), Xu et al. (2020).
[17] Chen, Xu e Lu (2018), Zhang et al. (2020).
[18] Celis et al. (2019).

Tabela 5 – Conjuntos de dados utilizados nos experimentos de avaliação *offline*.

Base de dados	#Usuários	#Itens	#Interações
Amazon Beauty	631.986	112.565	701.528
Amazon Books	1.008.954	206.710	2.437.999
Amazon Games	2.766.656	137.249	4.624.615
BestBuy	1.268.702	69.858	1.865.269
Delicious	1.867	69.198	437.593
Delicious-PU	1.867	38.576	93.188
MovieLens-100K	943	1.682	100.000
MovieLens-25M	162.541	59.047	25.000.095
RetailRocket	1.407.580	235.061	2.755.641

Figura 2 – Fluxo experimental adotado para a avaliação *offline* dos algoritmos lineares.



Fonte: Adaptado de Pires et al. (2025).

Organização temporal dos dados

Cada conjunto de dados foi ordenado de forma cronológica, simulando um ambiente sequencial em que as preferências do usuário evoluem ao longo do tempo. Essa ordenação permite aproximar o processo de recomendação da realidade, ainda que em um contexto sem interação efetiva.

Particionamento

O histórico de interações foi dividido em duas partes principais:

- **Treino (50% iniciais)**: utilizado como fase de *aquecimento* dos modelos. Dentro dessa fração, 10% foram separados para validação de hiperparâmetros.

- **Teste (50% finais)**: empregado para avaliar os algoritmos de forma incremental, em lotes sequenciais de 10% das interações. Essa divisão em lotes permite observar a evolução do desempenho ao longo do tempo, em vez de considerar apenas um único valor agregado.

Representação de usuários e itens

Para viabilizar a aprendizagem dos modelos lineares, foram geradas representações vetoriais (*embeddings*) dos itens a partir da fase de treino, utilizando o algoritmo *Alternating Least Squares* (ALS) (HU; KOREN; VOLINSKY, 2008). Os estados dos usuários foram calculados como a média dos vetores correspondentes aos itens previamente consumidos. Assim, cada interação usuário–item é representada por um par de vetores, que compõem o contexto do problema de bandits.

Treinamento incremental

Durante a fase de teste, a cada lote de interações, os agentes geraram recomendações, receberam o retorno observado (recompensa registrada nos dados históricos) e atualizaram seus modelos lineares de forma incremental. Esse processo busca aproximar a dinâmica de aprendizado contínuo típica dos sistemas de recomendação interativos.

Hiperparâmetros

Os hiperparâmetros dos algoritmos foram ajustados a partir da partição de validação. Para o ALS, foram considerados diferentes números de dimensões latentes (32, 64 e 128), parâmetros de regularização e número de iterações (HU; KOREN; VOLINSKY, 2008). Para os algoritmos de bandits, os parâmetros associados ao controle da exploração (como ϵ, α e ν) foram selecionados de acordo com o desempenho em *Normalized Discounted Cumulative Gain* (NDCG) em uma tarefa de recomendação com corte em 20 itens (NDCG@20).

Implementação

Toda a implementação foi realizada em `Python 3`, utilizando a biblioteca `implicit`[1] para a geração de *embeddings* e o framework `Mab2Rec` (KADIOĞLU; KLEYNHANS, 2024) para a execução dos algoritmos de bandits lineares.

Esse protocolo estabelece uma base metodológica clara para a comparação entre os algoritmos, permitindo identificar padrões de desempenho de cada estratégia em diferentes conjuntos de dados.

5.5 Métricas de avaliação

Para a avaliação *offline*, foram consideradas duas dimensões principais: *acurácia* e *diversidade*. A fundamentação teórica completa das métricas encontra-se na Seção 2.3, sendo aqui apresentada apenas a forma como foram aplicadas no experimento.

[1] <https://github.com/benfred/implicit>

Acurácia

A métrica de acurácia utilizada foi o *Normalized Discounted Cumulative Gain*(NDCG), na configuração $NDCG@20$. Essa versão considera apenas os 20 primeiros itens recomendados, refletindo cenários práticos em que o usuário interage com uma fração limitada da lista. O cálculo em cada lote do conjunto de teste permitiu acompanhar a evolução da acurácia ao longo do tempo (SHANI; GUNAWARDANA, 2011).

Diversidade

A diversidade foi avaliada por meio da métrica de **novidade**, que considera a frequência relativa dos itens no conjunto de dados. Valores mais altos de novidade indicam que os algoritmos tendem a recomendar itens menos populares, sugerindo maior exploração.

Síntese

A análise conjunta de $NDCG@20$ e novidade possibilita investigar o impacto do dilema exploração–aprofundamento. Enquanto a acurácia reflete a recuperação de itens relevantes com base no histórico, a diversidade evidencia o potencial de descoberta de novos conteúdos. Assim, a avaliação *offline* fornece uma primeira visão sobre como diferentes políticas equilibram precisão e exploração.

5.6 Resultados

A seguir são apresentados os resultados obtidos na avaliação *offline* dos algoritmos lineares. A análise foi organizada em três partes: acurácia, diversidade e seleção de hiperparâmetros. Esses resultados permitem compreender de que forma as diferentes estratégias de exploração se comportam sob o protocolo de *replay evaluation*, destacando as limitações desse tipo de avaliação.

5.6.1 Acurácia

A Figura 3 apresenta a evolução do NDCG@20 ao longo dos lotes de teste, enquanto a Tabela 6 resume os valores finais agregados.

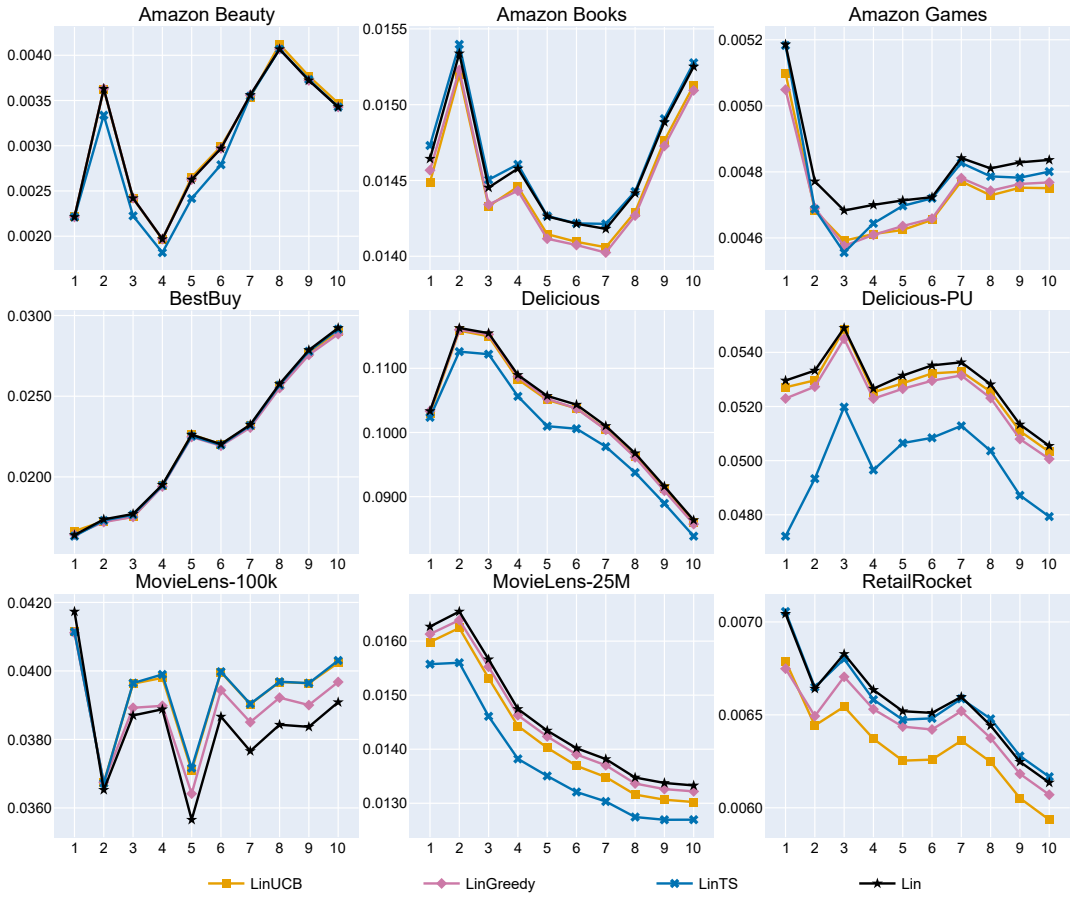
Observa-se que o desempenho das variantes exploratórias (LinGreedy, LinUCB e LinTS) acompanha de forma muito próxima o comportamento do modelo guloso (Lin). Em diversos casos, os algoritmos com exploração não conseguem se distinguir de maneira significativa, e em alguns cenários até apresentam resultados inferiores.

De forma agregada, o Lin obteve o melhor NDCG@20 em 56% dos conjuntos de dados e esteve entre os dois primeiros colocados em 89% dos casos. Em contrapartida, os algoritmos exploratórios superaram o modelo guloso apenas pontualmente, como no

MovieLens-100K, onde as características da base (catálogo reduzido e maior densidade de interações por usuário) favorecem a exploração.

Esses resultados confirmam que o protocolo *offline* tende a valorizar políticas que maximizam a exploração imediata dos itens já observados nos registros históricos, em detrimento de estratégias voltadas à exploração de novas opções.

Figura 3 – Evolução do NDCG@20 ao longo dos lotes de teste para cada conjunto de dados.



Fonte: autoria própria.

5.6.2 Diversidade

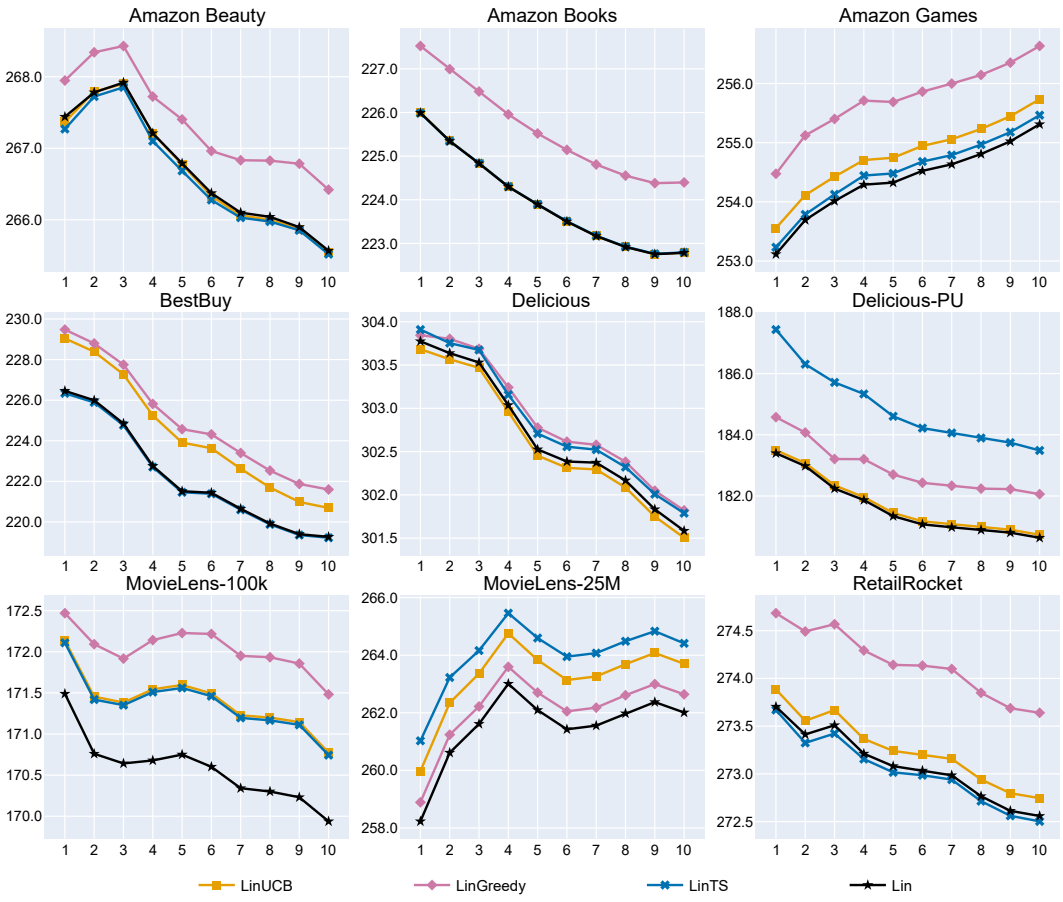
A Figura 4 apresenta os valores cumulativos de novidade ao longo dos lotes de teste. Como esperado, os algoritmos que incorporam mecanismos de exploração tendem a recomendar itens menos populares, resultando em maior novidade.

Entretanto, a análise revela uma relação inversa entre novidade e acurácia. Em bases como *Amazon Books* e *Amazon Games*, por exemplo, o aumento de diversidade observado

Tabela 6 – NDCG@20 agregado nos diferentes conjuntos de dados.

Dataset	**Lin**	**LinUCB**	**LinGreedy**	**LinTS**
Amazon Beauty	0,00342	**0,00346**	0,00342	0,00341
Amazon Books	0,01525	0,01512	0,01509	**0,01527**
Amazon Games	**0,00483**	0,00475	0,00477	0,00480
BestBuy	**0,02923**	0,02902	0,02885	0,02915
Delicious	**0,08634**	0,08596	0,08573	0,08385
Delicious-PU	**0,05055**	0,05033	0,05006	0,04794
MovieLens-100k	0,03900	0,04013	0,03958	**0,04021**
MovieLens-25M	**0,01332**	0,01302	0,01321	0,01269
RetailRocket	0,00614	0,00594	0,00607	**0,00617**

Figura 4 – Evolução da métrica de novidade ao longo dos lotes de teste.



Fonte: autoria própria.

no LinGreedy implicou em queda de NDCG@20. De forma semelhante, no *Delicious-PU* e no *MovieLens-25M*, o LinTS apresentou maior novidade, mas com acurácia inferior ao modelo guloso.

Esse comportamento reforça a limitação da avaliação *offline*: embora seja possível observar a geração de recomendações mais diversas, os ganhos em termos de descoberta não são devidamente refletidos nos registros históricos, levando a penalizações artificiais para estratégias exploratórias.

5.6.3 Seleção de hiperparâmetros

A seleção de hiperparâmetros foi conduzida na partição de validação, buscando maximizar o NDCG@20. A Tabela 7 resume os valores escolhidos para os parâmetros de controle de exploração de cada algoritmo.

Tabela 7 – Hiperparâmetros selecionados durante a validação para cada conjunto de dados.

Dataset	**LinUCB** (α)	**LinGreedy** (ϵ)	**LinTS** (ν^2)
Amazon Beauty	0,10	0,01	0,10
Amazon Books	0,10	0,01	0,50
Amazon Games	0,10	0,01	0,50
BestBuy	0,10	0,01	0,10
Delicious	0,10	0,01	0,50
Delicious-PU	0,10	0,01	1,00
MovieLens-100k	0,10	0,01	0,10
MovieLens-25M	0,10	0,01	1,00
RetailRocket	0,10	0,01	0,10

Os resultados mostram uma convergência consistente para valores que minimizam a exploração. Para o LinUCB e LinGreedy, por exemplo, os parâmetros α e ϵ foram quase sempre ajustados para os valores mais baixos da busca, próximos de zero. Já no caso do LinTS, houve alguma variação, mas mesmo quando valores mais altos foram selecionados, o desempenho em NDCG@20 permaneceu inferior ao do modelo guloso na maioria dos cenários.

Esse resultado reforça o viés intrínseco da avaliação *offline*: a própria etapa de otimização de hiperparâmetros tende a favorecer configurações que reduzem a exploração, o que compromete a avaliação justa de algoritmos projetados para equilibrar exploração e exploração.

5.7 Off-Policy Evaluation (OPE)

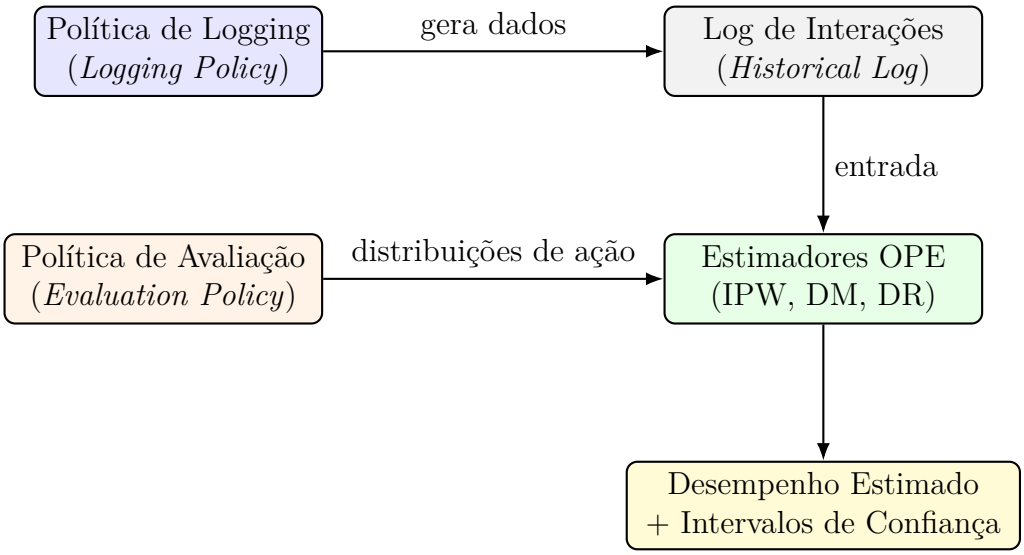
As análises anteriores mostraram que o protocolo de avaliação *offline* tende a favorecer políticas gulosas, penalizando algoritmos que incorporam mecanismos de exploração.

No *replay offline*, as recomendações feitas durante a avaliação não alteram os dados, que já estão fixados no *log* histórico. Isso significa que, quando um algoritmo sugere um item diferente daquele realmente exibido ou consumido no passado, não há como observar a recompensa que teria ocorrido caso essa recomendação alternativa tivesse sido feita. Na prática, essa restrição leva a uma avaliação enviesada e incompleta, pois apenas uma fração do espaço de decisões possíveis está representada nos dados (LI et al., 2011; DUDIK; LANGFORD; LI, 2011; WANG; AGARWAL; DUDÍK, 2017).

Para reduzir essa limitação, aplica-se a *Off-Policy Evaluation* (OPE). OPE é um conjunto de métodos estatísticos que permite estimar o desempenho de uma política de recomendação hipotética (a *política de avaliação*) utilizando dados coletados por outra política (a *política de logging*). Em outras palavras, busca-se responder: *como seria o desempenho de um algoritmo se ele tivesse sido usado para coletar os dados históricos, em vez da política que de fato os gerou?*

A Figura 5 ilustra o fluxo conceitual da avaliação *off-policy*. À esquerda, a política de *logging* corresponde ao algoritmo que originalmente gerou o *log* de interações observadas, composto por contextos, ações selecionadas (itens recomendados) e recompensas obtidas. Esse *log* histórico é então processado pelos estimadores de OPE, que combinam essas observações com as probabilidades de ação da política de avaliação, isto é, o algoritmo cujo desempenho se deseja estimar contrafactualmente. A partir dessa combinação, os estimadores (como IPW, DM e DR) ponderam as interações conforme o quanto as decisões da política avaliada divergem das decisões registradas pela política de *logging*. O resultado é uma estimativa do desempenho esperado da política avaliada, acompanhada de medidas de incerteza, como intervalos de confiança.

Figura 5 – Fluxo simplificado do processo de *off-policy evaluation* (OPE).



Fonte: autoria própria.

Neste trabalho, empregou-se o *Open Bandit Pipeline* (OBP) (SAITO et al., 2021), ferramenta de código aberto amplamente utilizada para avaliação contrafactual em *bandits* contextuais. O OBP disponibiliza um conjunto de dados real (*Open Bandit Dataset*) e implementações padronizadas de diferentes estimadores de OPE, favorecendo reprodutibilidade e comparabilidade.

5.7.1 Estimadores utilizados

Foram empregados três estimadores clássicos, que constituem a base da literatura de OPE:

- **Inverse Propensity Weighting (IPW)** (WANG; AGARWAL; DUDÍK, 2017): ajusta os resultados ponderando cada interação pela *propensão* (probabilidade de a ação ter sido escolhida pela política de logging). É não enviesado sob hipóteses adequadas, mas pode sofrer de alta variância quando as propensões são pequenas.

- **Direct Method (DM)** (DUDIK; LANGFORD; LI, 2011): aprende-se explicitamente um modelo de recompensa a partir dos dados históricos para estimar a recompensa esperada em qualquer par contexto–ação. Reduz variância, mas pode introduzir viés se o modelo estiver mal especificado.

- **Doubly Robust (DR)** (DUDIK; LANGFORD; LI, 2011; WANG; AGARWAL; DUDÍK, 2017): combina IPW e DM. É *duplamente robusto*: consistente se pelo menos um dos componentes (modelo de recompensa ou propensões) estiver corretamente especificado, equilibrando viés e variância na prática.

5.7.2 Implementação neste trabalho

O OBP foi utilizado como núcleo da avaliação, integrando as políticas lineares (*Lin*, *LinGreedy*, *LinUCB*, *LinTS*). Cada política treinada foi convertida em uma *distribuição de ações* (probabilidades de recomendar cada item por contexto), utilizada pelos estimadores do OBP para o cálculo contrafactual.

Para estimar incerteza, empregou-se bootstrap com 10 000 reamostragens e foram construídos intervalos de confiança de 95%. Assim, além do valor pontual estimado, avalia-se a variabilidade estatística associada a cada política.

5.8 Resultados com Off-Policy Evaluation (OPE)

Os resultados *off-policy* foram organizados em: (i) **valores absolutos estimados**; (ii) **valores relativos** versus uma política de referência (*Random*); e (iii) **intervalos de**

confiança de 95%. Essa apresentação permite comparar desempenhos médios e verificar significância estatística das diferenças.

A Tabela 8 apresenta os valores absolutos estimados pelos diferentes métodos de OPE (IPW, DM e DR). Observa-se que a política puramente gulosa (**Lin**) mantém os maiores valores em todos os casos. Esse efeito é especialmente visível nos estimadores que utilizam as propensões (IPW e DR). As propensões correspondem à probabilidade de cada ação ter sido selecionada pela política que gerou os dados históricos; portanto, servem como fator de correção do viés. Quando um item aparece nos *logs* com baixa probabilidade, mas seria escolhido com frequência pela nova política, sua contribuição deve ser amplificada. Da mesma forma, ações que já eram muito prováveis recebem menor peso. Assim, a dependência das propensões evidencia como a avaliação tende a favorecer políticas semelhantes à original, como é o caso do modelo guloso, reproduzindo o viés já observado nas análises *offline* (LI et al., 2011; DUDIK; LANGFORD; LI, 2011; WANG; AGARWAL; DUDÍK, 2017; SAITO et al., 2021).

Tabela 8 – Avaliação *off-policy* com IPW, DM e DR — valores absolutos estimados.

Estimador	**Lin**	**LinUCB**	**LinGreedy**	**LinTS**
IPW	**0,01637**	0,01270	0,01509	0,01176
DM	**0,00403**	0,00401	0,00383	0,00381
DR	**0,01623**	0,01499	0,01458	0,01182

A Tabela 9 apresenta os intervalos de confiança de 95% obtidos via *bootstrap* (10 000 reamostragens) para as estimativas *off-policy* com três métodos: IPW, DM e DR. Observa-se o *trade-off* clássico entre variância e viés:

- **IPW e DR** produzem intervalos de confiança mais largos, refletindo maior variabilidade nas estimativas. Em contrapartida, seus limites inferiores tendem a ser mais elevados, sugerindo menor viés e maior robustez na detecção de diferenças entre políticas.
- **DM** apresenta intervalos extremamente estreitos, evidenciando alta precisão estatística. Contudo, essa estabilidade depende fortemente da qualidade do modelo de recompensa utilizado, o que pode introduzir viés caso o modelo esteja mal especificado.

Além disso, a análise dos intervalos permite caracterizar o perfil de cada algoritmo. O método **Lin** frequentemente apresenta os maiores limites inferiores nos estimadores IPW e DR, indicando desempenho consistente e evidenciando o favorecimento de políticas gulosas na avaliação *off-policy*. Já o **LinUCB** e o **LinTS** exibem intervalos mais amplos e limites inferiores menores, refletindo maior incerteza, efeito esperado de algoritmos que exploram mais e, portanto, sofrem com menor sobreposição em relação à política de coleta. O **LinGreedy**, por sua vez, apresenta resultados intermediários, próximos aos de **Lin**,

Tabela 9 – Intervalos de confiança de 95% das estimativas *off-policy* (IPW, DM, DR). O melhor intervalo em cada linha (maior limite inferior) está destacado em negrito.

Estimador	**Lin**	**LinGreedy**	**LinUCB**	**LinTS**
IPW	**[0,0113; 0,0218]**	[0,0103; 0,0204]	[0,0085; 0,0173]	[0,0076; 0,0164]
DM	**[0,0040; 0,0041]**	[0,0040; 0,0041]	[0,0038; 0,0039]	[0,0038; 0,0039]
DR	**[0,0112; 0,0217]**	[0,0102; 0,0203]	[0,0104; 0,0192]	[0,0076; 0,0164]

mas sem alcançar os melhores limites inferiores em nenhum dos cenários analisados, o que reforça que a exploração puramente aleatória não trouxe vantagens na avaliação *offline*.

Assim, não há um estimador universalmente superior: a escolha depende do equilíbrio desejado entre robustez estatística (IPW/DR, com menor viés porém maior variância) e precisão (DM, com intervalos estreitos porém sujeitos a viés de modelagem). Essa constatação reforça a complexidade inerente à avaliação *off-policy* em CMAB.

Por fim, a Tabela 10 mostra os valores relativos em relação à política aleatória (*Random*). Valores maiores que 1 indicam desempenho superior à política originalmente adotada . Observa-se novamente a superioridade de *Lin* sob IPW/DR, enquanto DM atenua diferenças.

Tabela 10 – Valores relativos estimados em relação à política *Random*.

Estimador	**Lin**	**LinGreedy**	**LinUCB**	**LinTS**
IPW	**4,31**	3,97	3,34	3,10
DM	**1,06**	1,05	1,01	1,00
DR	**4,27**	3,95	3,84	3,11

Em síntese, mesmo com OPE, políticas exploratórias (*LinUCB*, *LinGreedy*, *LinTS*) não superam a política puramente gulosa (*Lin*) no cenário analisado. Isso reforça a conclusão de que a avaliação baseada em *logs* tende a favorecer estratégias próximas à política de coleta, limitando a capacidade de mensurar benefícios genuínos de exploração (LI et al., 2011; DUDIK; LANGFORD; LI, 2011; WANG; AGARWAL; DUDÍK, 2017; SAITO et al., 2021).

5.9 Considerações finais

A avaliação *offline*, apesar de sua ampla adoção na literatura e na prática industrial, mostrou-se limitada para mensurar de forma justa o desempenho de algoritmos de *bandits* contextuais lineares. Os resultados apresentados ao longo deste capítulo revelaram um padrão consistente: o modelo puramente guloso (*Lin*) superou ou igualou as variantes exploratórias em grande parte dos cenários analisados. Esse comportamento decorre de

um viés estrutural do protocolo de *replay evaluation*, que tende a valorizar escolhas coincidentes com as interações registradas no histórico, penalizando recomendações alternativas que não encontram respaldo nos dados disponíveis (LI et al., 2010; SILVA et al., 2022).

A análise de acurácia evidenciou que as curvas de desempenho dos algoritmos exploratórios seguem trajetórias muito próximas ao modelo guloso, raramente apresentando ganhos consistentes. Já na dimensão de diversidade, observou-se que políticas exploratórias de fato geram recomendações mais variadas e com maior grau de novidade. Contudo, tais ganhos não se traduzem em melhorias nas métricas de acurácia, reforçando a desconexão entre o potencial da exploração e o que é capturado pela avaliação baseada em registros históricos.

A etapa de seleção de hiperparâmetros reforçou esse viés: em praticamente todos os cenários, os melhores valores identificados foram aqueles que minimizam a exploração. Isso significa que o próprio processo de ajuste de parâmetros tende a convergir para configurações que favorecem o comportamento guloso, em detrimento de políticas mais exploratórias.

Mesmo com o emprego de técnicas de OPE, os resultados permaneceram alinhados à mesma tendência. Em particular, os estimadores projetados para reduzir o viés em relação à política de *logging* (como IPW e DR) ainda apontaram a política gulosa como dominante. Esses achados indicam que, embora a OPE represente um avanço metodológico em relação ao *replay* simples, ela ainda não resolve completamente o problema de subavaliação da exploração (DUDIK; LANGFORD; LI, 2011; LI et al., 2011; WANG; AGARWAL; DUDÍK, 2017).

Diante desse cenário, é possível retomar a PP1 (*Como avaliar de maneira justa CMAB com exploração adaptativa? Quais condições e protocolos são mais adequados?*) para concluir que, ao menos no escopo da avaliação *offline*, não foram encontradas condições capazes de oferecer respostas plenamente satisfatórias. Tanto o *replay evaluation* quanto a OPE apresentam limitações estruturais que levam à supervalorização de políticas gulosas e à subavaliação de estratégias exploratórias.

Esses achados reforçam que os métodos de avaliação *offline*, mesmo enriquecidos por OPE, não são suficientes para capturar o valor de políticas que adaptam a exploração ao longo do tempo. Tal limitação motiva a busca por protocolos alternativos, baseados em simuladores interativos, capazes de incorporar a dinâmica de múltiplas sessões, o tempo de retorno dos usuários e a possibilidade de recompensas contrafactuais. No próximo capítulo, será apresentada a avaliação *online* realizada com o simulador KuaiSim (ZHAO et al., 2023), adaptado para este trabalho, como uma alternativa mais realista e informativa para o estudo do dilema exploração–aprofundamento.

Capítulo 6

Simulação Online

Diante da limitação observada na avaliação *offline*, torna-se necessário adotar um protocolo alternativo que permita observar o desempenho das políticas em condições mais próximas de um ambiente interativo. A simulação online surge como uma solução viável para esse desafio, pois possibilita conduzir experimentos controlados em que tanto a escolha das recomendações quanto as respostas dos usuários podem ser modeladas de forma explícita. Ao contrário da avaliação *offline*, essa abordagem oferece flexibilidade para investigar o impacto de diferentes estratégias de exploração, além de permitir a incorporação de variáveis temporais que influenciam o comportamento do usuário, como o intervalo de tempo entre interações.

Neste capítulo, apresenta-se a utilização do *KuaiSim*, inicialmente proposto por Zhao et al. (2023) como simulador de referência, destacando sua adequação para experimentos que envolvem *bandits* contextuais sensíveis ao tempo. São descritas inicialmente as características e o funcionamento do simulador, seguidos das modificações realizadas para atender às necessidades específicas desta pesquisa, incluindo a adaptação para suportar políticas lineares e a modelagem de dinâmicas inter-sessões. Por fim, discute-se o protocolo experimental empregado, o qual viabilizou a comparação sistemática entre os métodos de referência e a abordagem proposta.

6.1 Motivação

A análise conduzida por meio de avaliação *offline* e métodos *off-policy* demonstrou limitações relevantes para o estudo de políticas de recomendação com exploração adaptativa. Em particular, verificou-se que tais protocolos tendem a privilegiar políticas determinísticas e pouco exploratórias, reproduzindo os vieses do conjunto de dados que as

originaram. Como consequência, o potencial de algoritmos que dependem de mecanismos de exploração para diversificar recomendações ou captar preferências latentes não pode ser avaliado de forma justa.

Essa limitação é especialmente crítica no contexto desta pesquisa, cujo foco está no uso da informação temporal — em especial o intervalo temporal entre interações — como modulador do nível de exploração. Para investigar adequadamente essa hipótese, é necessário um ambiente que permita observar não apenas a escolha imediata de itens, mas também o impacto dessas escolhas no retorno futuro do usuário, capturando a relação entre exploração, retenção e dinâmica temporal.

A simulação *online* atende a essa necessidade ao fornecer um ambiente controlado em que o ciclo completo de interação pode ser reproduzido: seleção da recomendação, resposta do usuário e evolução temporal do sistema. Tal abordagem permite que diferentes políticas sejam comparadas sob condições equivalentes, garantindo maior robustez na análise de desempenho. Além disso, a possibilidade de modelar explicitamente fatores temporais torna viável avaliar de forma sistemática como o tempo influencia a exploração em algoritmos de *bandits* contextuais.

Nesse sentido, a adoção de simulação *online* complementa as análises *offline*, oferecendo uma perspectiva mais realista e controlada sobre o comportamento dos algoritmos. Esse protocolo experimental é, portanto, essencial para sustentar a avaliação da proposta apresentada nesta pesquisa.

6.2 A base de dados KuaiRand

O *KuaiRand* é o conjunto de dados que serve como base para o simulador *KuaiSim*. Ele foi coletado a partir de interações reais em uma plataforma de vídeos curtos, com o objetivo de oferecer uma base pública e abrangente para pesquisa em recomendação sequencial. Seu diferencial em relação a outros conjuntos reside na presença de informação temporal detalhada, permitindo modelar resposta imediata (por exemplo, clique) e também retenção entre sessões (retorno do usuário após certo intervalo). Essa combinação viabiliza estudos que integram engajamento dentro da sessão e dinâmica inter-sessões em um ambiente único de simulação.

6.2.1 Características gerais

O *KuaiRand* contém três componentes principais:

1. ***Log* de interações**, que registra usuário (`user_id`), item (`video_id`), carimbo temporal (`time_ms`) e resposta imediata (`is_click`);

2. **Atributos de usuários**, compostos por variáveis categóricas e numéricas;

3. **Atributos de itens** (vídeos), que incluem informações de conteúdo.

Esses dados permitem construir variáveis de contexto para algoritmos de *bandits* contextuais e calcular o intervalo entre interações consecutivas de um mesmo usuário (δ_t).

Retenção e padrão de retorno

Um aspecto importante observado na análise exploratória é o comportamento da retenção dos usuários, ou seja, a probabilidade de que um usuário volte a interagir após certo número de dias. Para cada usuário, calculou-se o intervalo em dias entre interações consecutivas, e a distribuição agregada desses intervalos mostrou um padrão bastante característico: a maioria dos retornos acontece no dia seguinte ($\delta_t = 1$), e a frequência de retorno diminui de forma rápida à medida que o intervalo aumenta. Em termos práticos, isso significa que a maior parte dos usuários retorna em curtos períodos, enquanto apenas uma parcela cada vez menor volta após vários dias.

Esse padrão de decaimento é semelhante ao de uma distribuição geométrica, na qual a probabilidade de retorno cai aproximadamente de forma exponencial com o tempo. Esse achado é relevante porque reflete o comportamento típico em plataformas de consumo rápido de conteúdo: usuários altamente ativos retornam em intervalos muito curtos, enquanto outros apresentam pausas mais longas entre as sessões. Para o *KuaiSim*, essa característica temporal é central, pois permite simular de forma mais realista a dinâmica de engajamento e retenção dos usuários.

Verificação da hipótese geométrica (PMF e similaridade JS)

Para verificar a compatibilidade do padrão observado de retornos com uma distribuição geométrica, foi construída a função de massa de probabilidade (PMF) empírica dos intervalos δ_t e, em seguida, foi sobreposta a PMF de uma distribuição geométrica ajustada aos dados (Figura 6). O ajuste foi realizado pelo estimador de máxima verossimilhança do parâmetro p da geométrica com suporte discreto $\{1, 2, \dots\}$, dado por

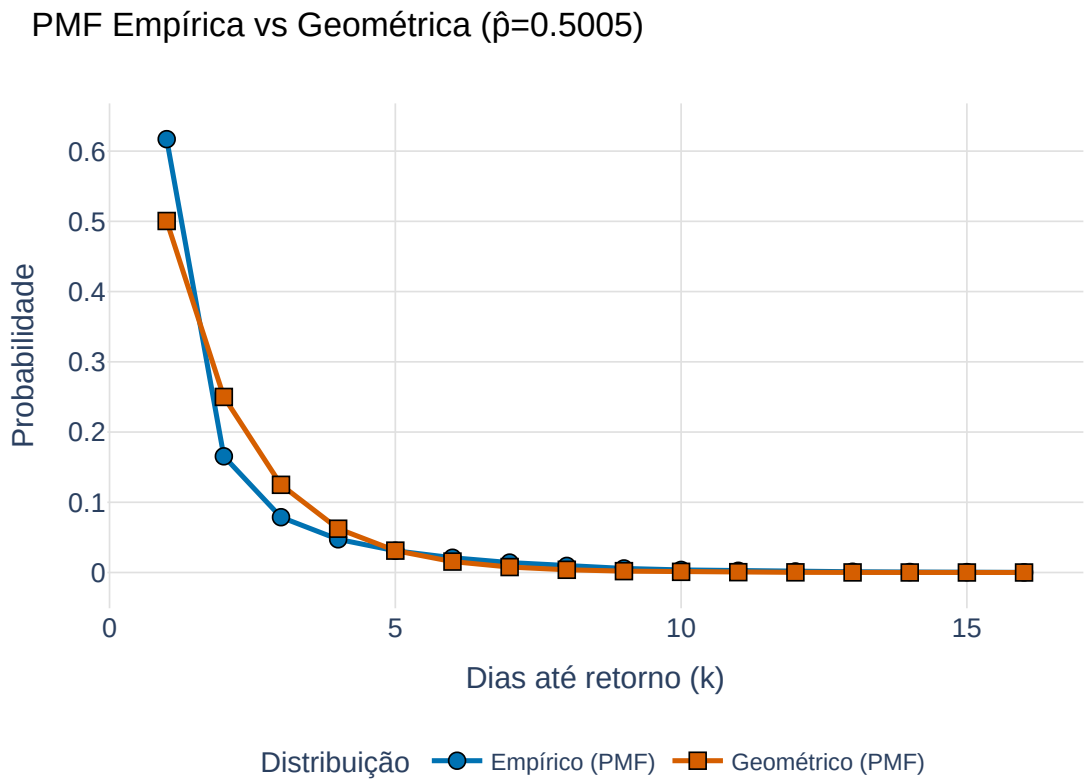
$$\hat{p} \;=\; \frac{1}{\bar{x}}, \qquad \Pr(X = k) \;=\; \hat{p}\,(1-\hat{p})^{k-1}, \;\; k = 1, 2, \dots$$

em que $\bar{x}$ denota a média empírica dos intervalos. No conjunto analisado, obteve-se $\hat{p} = 0{,}500508$, o que implica retorno médio de aproximadamente $1/\hat{p} \approx 2$ dias.

A proximidade entre as distribuições foi quantificada por meio da similaridade de Jensen–Shannon (JS), que mede a sobreposição de duas distribuições de probabilidade. A divergência de Jensen–Shannon é definida como

$$JS(P \parallel Q) \;=\; \tfrac{1}{2}\,KL(P \parallel M) \;+\; \tfrac{1}{2}\,KL(Q \parallel M), \qquad M = \tfrac{1}{2}(P+Q),$$

Figura 6 – PMF empírica dos intervalos de retorno sobreposta à PMF geométrica ajustada.



Fonte: autoria própria.

em que KL representa a divergência de Kullback–Leibler. Para fins de interpretação, foi utilizada a forma normalizada da similaridade,

$$S_{\mathrm{JS}} \;=\; 1 - \frac{JS(P \parallel Q)}{\ln 2},$$

de modo que $S_{\mathrm{JS}} = 1$ indica distribuições idênticas e valores próximos de 0 indicam distribuições muito diferentes. No caso estudado, foi obtido $S_{\mathrm{JS}} = \mathbf{97{,}99}\%$, indicando forte aderência da forma geométrica ao padrão empírico.

Do ponto de vista visual (Figura 6), a curva geométrica reproduz o pico de retornos em $\delta_t = 1$ e o decaimento subsequente, com discrepâncias residuais na cauda. Assim, a hipótese geométrica apresenta-se como aproximação adequada para descrever o comportamento agregado de retorno, fornecendo sustentação quantitativa à premissa adotada no *KuaiSim*.

6.2.2 Pré-processamento da base de dados

Antes de ser utilizado no simulador, o *KuaiRand* passa por um pré-processamento destinado a reduzir a esparsidade, estruturar sessões de interação e tratar valores ausentes em

atributos de usuários e itens. Esse procedimento garante que o ambiente simulado reflita um cenário realista e, ao mesmo tempo, computacionalmente manejável. As principais etapas foram:

1. **Filtragem de usuários e itens pouco ativos**: a base de dados original contém mais de 27 mil usuários, 7,5 mil vídeos e aproximadamente 1,4 milhão de registros de interação. Para reduzir ruídos associados a usuários ou itens com poucas interações, foi aplicada auma função que remove amostras de baixa frequência. Após a filtragem, o conjunto resultante conta com cerca de 19,5 mil usuários, 5,6 mil itens e 1,34 milhão de registros, mantendo a maior parte da densidade original de interações.

2. **Construção de sessões e posições**: cada interação foi agrupada em sessões diárias por usuário, de modo que todas as ações de um usuário em uma mesma data constituem uma sessão. Além disso, foi criado um índice de posição para identificar a ordem em que os itens foram consumidos dentro da sessão. Essa informação é crucial para simulações que reproduzem a sequência de recomendações e respostas do usuário.

3. **Conversão e padronização de datas**: o campo temporal original (`time_ms`, em milissegundos) foi convertido para o formato de data (`YYYYMMDD`), permitindo análises no nível de dia e viabilizando o cálculo de intervalos de retorno entre sessões.

4. **Tratamento dos atributos de itens**: os dados contém atributos de cada vídeo, como categorias e tipo de música. Valores ausentes nessas colunas foram preenchidos com zero.

5. **Tratamento dos atributos de usuários**: os atributos de usuário incluem diversas codificações (*one-hot encoding*). Para evitar inconsistências durante o treinamento e simulação, valores ausentes nessas colunas foram preenchidos com o valor `-1`, representando categoria indefinida.

Ao final do pré-processamento, o *KuaiRand* foi transformado em um conjunto mais compacto, estruturado por sessões e com atributos devidamente tratados, pronto para ser integrado ao *KuaiSim*.

6.3 O simulador KuaiSim

O *KuaiSim* é um simulador desenvolvido recentemente com o objetivo de fornecer um ambiente abrangente para avaliação de sistemas de recomendação baseados em aprendizado por reforço (ZHAO et al., 2023). Sua construção foi motivada pelas limitações

observadas em simuladores anteriores, que frequentemente assumiam respostas simplificadas do usuário, ignoravam sinais de retenção entre sessões ou eram restritos a cenários muito específicos. Para superar tais restrições, o *KuaiSim* foi projetado a partir de dados reais do conjunto *KuaiRand*, abrangendo múltiplos tipos de *feedback* e permitindo a modelagem de interações intra e inter-sessões.

O funcionamento do simulador organiza-se em três níveis de tarefas, que correspondem a diferentes perspectivas de avaliação em recomendação:

- **Nível de requisição (*request-level*):** considera uma única interação em que uma lista de itens é exibida ao usuário. Neste nível, o foco está em capturar as correlações entre itens dentro da lista, avaliando o impacto conjunto das recomendações por meio de métricas como recompensa média, cobertura e diversidade intra-lista (ILD).

- **Nível de sessão (*whole-session*):** corresponde ao ciclo completo de interações dentro de uma mesma sessão, desde o momento em que o usuário inicia até o ponto em que decide encerrar sua utilização. Nesse caso, cada recomendação influencia o estado do usuário e afeta o desempenho acumulado ao longo da sessão.

- **Nível inter-sessões (*cross-session*):** expande a análise para múltiplas sessões de um mesmo usuário, incorporando o conceito de tempo de retorno. Esse nível permite investigar a retenção de usuários e avaliar como a qualidade das recomendações influencia o intervalo entre sessões consecutivas.

Para viabilizar essa modelagem, o *KuaiSim* integra três módulos principais que representam dimensões complementares do comportamento do usuário: **resposta imediata**, **saída da sessão** e **retenção**. O primeiro trata da reação instantânea ao conteúdo exibido; o segundo modela a decisão de encerrar uma sessão em andamento; e o terceiro estima o intervalo até o retorno do usuário em uma nova sessão. Juntos, esses módulos permitem que o simulador conecte *decisões locais* (em nível de item e de sessão) a *efeitos globais* (retenção e engajamento de longo prazo).

A implementação do *KuaiSim* baseia-se em modelos de usuário pré-treinados com dados de *log* do *KuaiRand*, garantindo maior consistência com padrões reais de comportamento. O repositório disponibiliza procedimentos de preparação de dados, treinamento de modelos de resposta e *scripts* de execução de experimentos nos diferentes níveis de tarefa. Além disso, oferece protocolos de avaliação padronizados e algoritmos de referência, permitindo estabelecer comparações justas entre diferentes abordagens.

Em comparação com simuladores existentes, como o *RecSim* (IE et al., 2019), o *RecoGym* (ROHDE et al., 2018), o *Virtual-Taobao* (SHI et al., 2019) e o *RL4RS* (WANG et al., 2023), o *KuaiSim* apresenta uma cobertura mais ampla. Enquanto os primeiros oferecem suporte limitado a um subconjunto de tarefas (geralmente focados apenas em *feedback* imediato ou sessões únicas), o *KuaiSim* é o único que integra de forma consistente

os três níveis de interação: requisição, sessão e inter-sessões. Essa abrangência, combinada com o uso de dados imparciais do *KuaiRand*, confere ao simulador maior realismo e flexibilidade para o estudo de algoritmos de recomendação sensíveis ao tempo.

6.4 Funcionamento interno do KuaiSim

O *KuaiSim* distingue-se de simuladores anteriores por não adotar regras fixas e artificiais para gerar interações, mas por aprender a simular o comportamento do usuário a partir de dados reais. Para isso, utiliza o conjunto *KuaiRand*, que fornece interações entre usuários e vídeos com múltiplos tipos de *feedback*. A partir desses registros, o simulador treina modelos neurais capazes de reproduzir padrões observados no mundo real, garantindo maior realismo no processo de simulação.

O núcleo dessa arquitetura está na implementação dos três módulos centrais: resposta imediata, saída da sessão e retenção. A seguir, cada módulo é detalhado separadamente, mas sempre mantendo a conexão entre eles, de modo a mostrar como, em conjunto, constroem uma simulação integrada das jornadas de interação do usuário.

6.4.1 Módulo de resposta imediata

O núcleo do simulador é o modelo de resposta do usuário. Esse modelo supervisionado recebe como entrada:

- ❏ características do usuário (identificador e atributos de perfil);
- ❏ características dos items recomendados;
- ❏ histórico das interações anteriores do usuário.

Essas informações são convertidas em representações vetoriais (*embeddings*), que permitem ao modelo capturar similaridades e padrões latentes. Em seguida:

1. O histórico é processado por um codificador sequencial (*Transformer*), que considera tanto a ordem dos eventos quanto o tipo de *feedback* associado a cada interação.
2. O vetor do usuário e o vetor do histórico são concatenados, formando o estado do usuário.
3. Esse estado é combinado com a representação do item candidato em uma rede neural (*scorer*), que estima a probabilidade de ocorrência de cada tipo de *feedback*.

Assim, para cada item exibido, o modelo retorna a probabilidade de eventos como clique, curtida, comentário, compartilhamento, seguir, rejeição ou visualização prolongada.

Pseudocódigo do Modelo de Resposta

O funcionamento pode ser resumido no seguinte pseudocódigo de alto nível:

Algoritmo 1 Cálculo da probabilidade de feedbacks

Entrada: interação (u, i, H) com usuário u, item i e histórico H
Saída: vetor de probabilidades por tipo de feedback
Codificar usuário $u \rightarrow$ vetor U
Codificar item $i \rightarrow$ vetor I
Codificar histórico $H \rightarrow$ vetor H_{enc}
Construir estado do usuário $S \leftarrow \text{concat}(U, H_{enc})$
for cada tipo de feedback f **do**
    $probabilidade[f] \leftarrow scorer(S, I)$
end for
return *probabilidade*

Essas probabilidades são utilizadas pelo simulador para amostrar respostas, de forma que cada execução do mesmo cenário pode resultar em diferentes trajetórias, refletindo a natureza estocástica das interações reais.

6.4.2 Módulo de saída da sessão

Esse módulo complementa o anterior ao determinar a duração da sessão. Enquanto o módulo de resposta imediata define os sinais obtidos em cada recomendação, o de saída traduz o acúmulo desses sinais em uma decisão de continuar ou abandonar a sessão.

Diferentemente de um critério arbitrário por número fixo de passos, o *KuaiSim* utiliza uma variável latente de paciência/satisfação (*temper*) que é atualizada a cada requisição com base nos *feedbacks* observados. A ideia é que recomendações pouco satisfatórias aumentam a probabilidade de saída do usuário; já recomendações boas desaceleram (ou não aceleram) essa saída.

Em cada passo, o simulador:

1. coleta os *feedbacks* imediatos do usuário para o *slate* (conjunto de itens) recomendado (clique, *like*, *hate*, etc.);

2. converte os *feedbacks* do *slate* em recompensas por item, ponderadas por pesos por tipo de *feedback* (*response_weights*);

3. agrega por item (soma nos tipos de *feedback*) e depois por *slate* (média nos itens) para obter um reforço médio do passo;

4. atualiza a variável *temper* com base nesse reforço médio, com saturação em um intervalo negativo;

5. sinaliza `done` (saída) quando $temper < 1$.

Formulação

Seja $R_{t,i}$ a recompensa do i-ésimo item do *slate* no passo t, obtida ao ponderar os *feedbacks* binários pelos respectivos pesos; seja $\bar{r}_t = \frac{1}{L}\sum_{i=1}^{L} R_{t,i}$ a média no *slate* de tamanho L. Define-se a atualização:

$$\Delta_t \;=\; \mathrm{clip}\Big(\,\bar{r}_t - c\,,\,[-2,0]\,\Big), \qquad \mathrm{temper}_{t+1} \;=\; \mathrm{temper}_t + \Delta_t,$$

com $c = 2$, "clip$(\cdot,[-2,0])$" limitando a atualização ao intervalo $[-2,0]$ (incrementos positivos são truncados a 0, e decréscimos excessivos são limitados a -2). O usuário deixa a sessão quando

$$\mathrm{temper}_{t+1} \;<\; 1.$$

Os hiperparâmetros relevantes (valor inicial $\mathrm{temper}_0 =$ `initial_temper`, limiar de saída 1 e a constante c) controlam o horizonte médio da sessão e a sensibilidade à qualidade das recomendações. Esse desenho operacionaliza a noção de que *feedbacks* imediatos afetam a decisão de saída.

Pseudocódigo (alto nível)

Algoritmo 2 Atualização de paciência do usuário

Entrada: resposta imediata do *slate* $R[B, L, n_f]$; pesos $w[n_f]$
Estado: vetor de paciência $temper[B]$
Saída: máscara $done[B]$
for $b \leftarrow 1$ **to** B **do**
    $r \leftarrow$ recompensa por item ponderada pelos tipos de feedback
    $m \leftarrow$ média de r no *slate*
    $temper_b \leftarrow temper_b + \mathrm{clip}(m - c,\, -2,\, 0)$
    $done_b \leftarrow (temper_b < 1)$
end for

Observações

- **Interpretação do corte em c:** somente baixa recompensa média acelera a saída; ganhos marginais não "recarregam" a paciência (incremento positivo é limitado a 0). Isso torna o tempo de sessão mais realista e robusto a ruídos.
- **Agregação em *slate*:** a decisão de saída considera o efeito conjunto dos itens expostos (média por *slate*).
- **Hiperparâmetros:** `initial_temper` (valor padrão 10), `max_step_per_episode` (100), c (2) e os limites da função *clip* (intervalo $[-2,0]$) governam o comprimento médio da sessão e podem ser ajustados para alinhar com distribuições empíricas (profundidade).

6.4.3 Módulo de retenção (estimativa de δ_t)

Por fim, o módulo de retenção fecha o ciclo de interação ao estimar o intervalo entre sessões consecutivas. Se a resposta imediata reflete a reação instantânea e o módulo de saída captura a profundidade de uma sessão, o de retenção expande a análise para efeitos de longo prazo, conectando recompensas recebidas a tempos de retorno.

O objetivo é estimar o tempo de retorno do usuário após o término de uma sessão, isto é, o intervalo δ_t entre a última requisição da sessão atual e a primeira requisição da próxima sessão. No *KuaiSim*, esse componente fecha o ciclo inter-sessões e permite avaliar efeitos de longo prazo, como retenção e cadência de uso.

Após cada sessão, a probabilidade do usuário voltar em d dias depende *(i)* do seu *perfil/propensão* individual e *(ii)* da *qualidade* (recompensas) das recomendações recebidas na sessão que acabou de terminar. O simulador modela essa dinâmica como uma distribuição geométrica parametrizada por uma probabilidade de retorno no dia seguinte p_{ret} que combina um viés global, um viés pessoal e um termo que responde à recompensa recente, e o retorno é então amostrado dessa distribuição. Além disso, a própria análise de dados do KuaiRand mostra que a distribuição empírica dos tempos de retorno tem formato geométrico, decaindo rapidamente após poucos dias.

Implementação

Na implementação, o módulo de retenção é um classificador sequencial que:

1. **Recebe** como entrada um *histórico de sessões* do usuário (codificações de sessão) e, opcionalmente, a sequência de *gaps* (dias) entre sessões anteriores;
2. **Codifica** a sequência com um *Transformer* para capturar dependências temporais e tendência de engajamento ao longo do tempo;
3. **Produz** uma distribuição discreta sobre os possíveis dias de retorno $\{1, \ldots, D\}$ por meio de uma rede neural profunda de saída (*softmax*);
4. **Treina** com *Cross-Entropy* (alvo: dia real de retorno observado nos *logs*).

Neste contexto, uma rede neural profunda (Deep Neural Network, DNN) é composta por múltiplas camadas de processamento não linear, capaz de modelar relações complexas entre variáveis de entrada e saída. Já a *Cross-Entropy* é uma função de perda amplamente utilizada em tarefas de classificação, que mede a discrepância entre a distribuição predita pelo modelo (no caso, via *softmax*) e a distribuição real observada nos dados (dia efetivo de retorno).

Durante o treino, o modelo observa sequências de sessões e o dia real de retorno e ajusta seus parâmetros minimizando a *Cross-Entropy*. Durante a simulação, dado o estado pós-sessão, o modelo produz uma distribuição $P(d{=}1), \ldots, P(d{=}D)$; o simulador

então amostra um dia de retorno. Essa amostragem introduz estocasticidade realista na cadência inter-sessões.

Pseudocódigo (alto nível)

Algoritmo 3 Predição do dia de retorno do usuário

Entrada: histórico de sessões $S = [s_1, \ldots, s_T]$; gaps anteriores $G = [g_1, \ldots, g_T]$
Saída: distribuição $P(d = 1..D)$ sobre o dia de retorno
$E_{\text{sess}} \leftarrow$ projeta(s_t) $\forall s_t \in S$
$E_{\text{pos}} \leftarrow$ embedding de posição temporal
$E_{\text{gap}} \leftarrow$ embedding do gap anterior (dias)
$Seq \leftarrow$ concat$(E_{\text{sess}} + E_{\text{pos}},\ E_{\text{gap}})$
$H \leftarrow$ Transformer(Seq)
$z \leftarrow$ normaliza(vetor_final(H))
$logits \leftarrow$ DNN(z)
$P \leftarrow$ softmax$(logits)$ {Distribuição sobre $\{1..D\}$}
$dia \leftarrow$ amostra(P) {Dia de retorno estimado}
return P, dia

Ligação com δ_t na avaliação

O valor amostrado d torna-se o δ_t que separa o fim de uma sessão do início da próxima. Isso permite:

1. mensurar retorno médio (dias) e taxa de retenção (probabilidade de retornar até D dias);
2. analisar como políticas mais ou menos exploratórias afetam δ_t via recompensa;
3. estudar decisões de curto prazo (recompensa imediata) e efeitos de longo prazo (cadência de retorno).

6.5 Métricas de avaliação

Na avaliação em ambiente de simulação online, foram selecionadas seis métricas principais: **recompensa média**, **cobertura**, **área sob a curva de cobertura (Coverage-AUC)**, **diversidade intra-lista (ILD)**, **tamanho médio da sessão** e **tempo médio de retorno**. A formulação detalhada dessas métricas encontra-se na Seção 2.3; aqui descreve-se apenas o papel de cada uma na análise dos experimentos.

- **Recompensa média**: métrica primária de eficácia, reflete a utilidade imediata das recomendações a partir das interações registradas.

- **Cobertura**: avalia a proporção do catálogo efetivamente exposta aos usuários, indicando o grau de exploração dos itens.
- **CoverageAUC**: considera não apenas a cobertura final, mas também a velocidade com que os itens passam a ser explorados ao longo da simulação.
- **ILD**: mede a variedade de itens dentro de cada lista de recomendação, evitando redundância e promovendo diversidade.
- **Tamanho médio da sessão**: indica o nível de engajamento, por meio da quantidade média de interações realizadas em cada sessão.
- **Tempo médio de retorno**: captura a dimensão temporal da retenção, medindo o intervalo médio entre sessões consecutivas de um mesmo usuário.

Essas métricas, em conjunto, permitem avaliar não apenas a qualidade imediata das recomendações, mas também a diversidade oferecida, a cobertura do catálogo e o impacto das políticas na cadência de interação e retenção dos usuários.

6.6 Modificações no simulador

Esta seção descreve as modificações realizadas no *KuaiSim* para suportar avaliação justa de *bandits* contextuais com dinâmica inter-sessões. O ambiente proposto, denominado *TimeAwareEnvironment*, unifica os módulos originalmente separados de *whole session* (resposta imediata) e *cross-session* (retenção/retorno), tornando o intervalo temporal entre sessões (δ_t) um sinal endógeno do ciclo de simulação e parte explícita do contexto observado pelos agentes.

6.6.1 Da versão original ao ambiente modificado

Em sua versão original, o módulo *whole-session* modela apenas a interação intra-sessão, com saída do usuário controlada por um parâmetro de *temper* e substituição imediata do usuário que saía. O módulo *cross-session* adiciona uma modelagem de retenção que amostra o dia de retorno, sendo este valor utilizado apenas para registro, sem alimentar o contexto da próxima sessão do mesmo usuário.

Como principal diferencial, o ambiente *TimeAwareEnvironment* integra ambos: quando os usuários encerram uma sessão, um modelo de retenção amostra o intervalo de retorno (δ_t). Em seguida, o ambiente reinicia a sessão do mesmo usuário, expondo δ_t (e sua média histórica δ_m) no contexto subsequente. Essa unificação permite que políticas contextuais utilizem diretamente informações temporais para modular sua exploração.

6.6.2 Arquitetura e componentes

Diferentemente do *KuaiSim* original, em que os processos de resposta imediata e de retorno inter-sessões eram tratados por módulos distintos, a modificação proposta nesta dissertação unifica esses elementos em um único ambiente integrado. Essa unificação permite que o intervalo entre sessões (δ_t) seja gerado de forma endógena, como consequência direta das interações do usuário com o sistema, em vez de ser tratado como uma variável externa.

O *TimeAwareEnvironment* organiza-se em três módulos acoplados:

1. **Resposta imediata** (*immediate response*): dado um *slate* recomendado, um modelo pré-treinado gera probabilidades de múltiplos *feedbacks* por item (por exemplo, cliques).

2. **Saída de sessão** (*leave*): a variável `temper` é atualizada a partir dos *feedbacks*; o usuário encerra a sessão quando o limiar de engajamento é atingido.

3. **Retenção/retorno** (*retention*): ao término da sessão, um modelo paramétrico amostra o tempo de retorno δ_t (via distribuição geométrica enviesada pelo estado do usuário e suas respostas), atualiza as estatísticas do histórico e inicia a próxima sessão do mesmo usuário.

Essa integração garante que os três processos façam parte de um mesmo ciclo de simulação, fornecendo aos agentes um contexto mais realista e sensível ao histórico de interações.

6.6.3 Interface para *bandits* contextuais

A interação entre agente e ambiente segue um formato direto: a cada passo, o agente escolhe quais itens recomendar, e o ambiente devolve duas informações principais:

- **Observação**: descreve o estado atual do usuário, incluindo seu perfil, o histórico de interações e os sinais temporais (δ_t, δ_m). O valor δ_t representa o intervalo de tempo até o retorno do usuário, enquanto δ_m guarda a média desses intervalos ao longo das sessões.

- **Feedback**: mostra como o usuário reagiu à recomendação, indicando cliques e outros tipos de resposta, se a sessão terminou e qual foi a recompensa resultante. Quando uma sessão acaba, o ambiente também informa o valor de δ_t daquele retorno.

Nos ambientes originais do KuaiSim, a interface seguia um formato inspirado em *reinforcement learning* (RL). Nessa configuração, além do perfil e do histórico do usuário, o ambiente fornecia diversas informações adicionais, como métricas de diversidade da lista

recomendada e até probabilidades de retorno em dias futuros. Em algumas variações, o agente não escolhia diretamente os itens, mas apenas pesos para combinar diferentes sinais de feedback, ficando a cargo do ambiente gerar a lista final de recomendações. Esse desenho é útil para cenários mais gerais de RL, mas traz uma complexidade desnecessária quando o objetivo é avaliar algoritmos de *bandits* contextuais lineares.

Para este trabalho, a interface foi modificada para ser mais clara e prática. O ambiente passou a entregar apenas o que é essencial para os algoritmos de *bandit*: o estado do usuário com seus sinais temporais e a resposta imediata às recomendações. A ação do agente também foi simplificada, de forma que ele escolhe diretamente quais itens recomendar, sem depender de mecanismos intermediários de ranqueamento. Outra mudança importante é que o próprio ambiente já calcula uma recompensa escalar que combina cliques e tempo de retorno, permitindo que os algoritmos utilizem esse valor de forma direta e comparável.

Essas modificações trazem três vantagens principais: reduzem a complexidade da interface, tornam a comparação entre diferentes políticas exploratórias mais justa e colocam a dimensão temporal no centro do contexto. Os valores de δ_t (intervalo até o retorno) e δ_m (média histórica dos intervalos de um mesmo usuário) são fornecidos de maneira explícita, o que abre espaço para algoritmos como o *Time-Aware LinBoltzmann* explorarem esses sinais de forma mais efetiva.

6.6.4 Modelagem temporal e uso de δ_t

O intervalo entre sessões, representado por δ_t, é definido por um modelo de retenção que considera três fatores: a tendência individual de cada usuário (seu viés pessoal), a influência dos *feedbacks* mais recentes e um viés global que privilegia retornos rápidos, como o do dia seguinte.

Quando o usuário retorna, esse intervalo é incorporado ao contexto fornecido ao agente, juntamente com uma estimativa da média histórica de retornos para aquele perfil. Dessa forma, os algoritmos passam a ter acesso não apenas ao estado imediato da interação, mas também a sinais temporais que refletem o ritmo de engajamento do usuário.

Políticas sensíveis ao tempo utilizam essas informações para calibrar o nível de exploração em suas escolhas, ajustando a probabilidade de recomendar itens menos conhecidos conforme a cadência de retorno do usuário.

6.6.5 Recompensas

O ambiente provê uma função de recompensa híbrida, construída para refletir tanto sinais de engajamento imediato quanto aspectos de retenção inter-sessões. A definição baseia-se em duas componentes complementares:

1. **Componente de engajamento imediato.** Cada interação item–usuário em um *slate* pode gerar múltiplos tipos de *feedback*, representados como variáveis binárias ($r_{u,a,f} \in \{0, 1\}$) que indicam se o evento ocorreu ou não para o item a recomendado ao usuário u. No conjunto *KuaiRand*, estão disponíveis sete tipos de sinais:

   - *is_click* — clique no item;
   - *long_view* — visualização prolongada;
   - *is_like* — marcação de "curtir";
   - *is_comment* — comentário publicado;
   - *is_forward* — compartilhamento do conteúdo;
   - *is_follow* — seguir o criador do conteúdo;
   - *is_hate* — feedback negativo explícito.

   Como a frequência desses sinais varia bastante no conjunto de dados, utiliza-se um vetor de **pesos** $\mathbf{w} \in \mathbb{R}^7$ para calibrar sua importância relativa. Esses pesos não são probabilidades diretas, mas sim valores que expressam a razão entre o número de ocorrências e não ocorrências de cada tipo de *feedback* no *log* de interações. Formalmente, o peso de um feedback f é dado por:

$$w_f \;=\; \frac{\#\{y_f = 1\}}{\#\{y_f = 0\}},$$

   em que $\#\{y_f = 1\}$ representa quantas vezes o evento f ocorreu e $\#\{y_f = 0\}$ quantas vezes ele não ocorreu.

   Dessa forma:

   - Se um tipo de feedback é comum (por exemplo, *is_click* ou *long_view*), o peso w_f será relativamente alto, refletindo sua importância no cálculo da recompensa.
   - Se o feedback é raro (como *is_comment* ou *is_forward*), o peso será muito próximo de zero.
   - Para *is_hate*, o peso calculado é multiplicado por -1, de modo que esse sinal contribua negativamente na recompensa.

   Em outras palavras, os pesos funcionam como fatores de ponderação: quanto mais frequente e relevante o sinal no conjunto de dados, maior a sua contribuição na recompensa de engajamento.

   Assim, para um usuário u e um *slate* de tamanho L, a recompensa de engajamento imediato é definida como:

$$R_u^{\text{feedback}} \;=\; \frac{1}{L} \sum_{a=1}^{L} \sum_{f=1}^{7} r_{u,a,f} \cdot w_f,$$

em que $r_{u,a,f}$ indica a ocorrência do *feedback* f no item a e w_f representa seu peso correspondente.

Tabela 11 – Pesos de cada tipo de *feedback* no conjunto KuaiRand.

Feedback	**Descrição**	**Peso** w_f
is_click	Clique no item recomendado	0.846
long_view	Visualização prolongada	0.492
is_like	Marcação de "curtir"	0.019
is_comment	Publicação de comentário	0.003
is_forward	Compartilhamento	0.001
is_follow	Novo "follow" no criador	0.001
is_hate	Feedback negativo explícito	−0.0005

2. **Componente de retenção.** Quando a sessão se encerra, o modelo de retenção amostra o intervalo de retorno δ_t (número de dias até a próxima sessão). Esse valor é incorporado na recompensa como $R_u^{\text{ret}} = 1/\delta_t$, de modo que retornos mais rápidos (menor δ_t) geram maior contribuição positiva.

A recompensa final é composta como combinação linear ponderada:

$$R_u = w_{\text{feedback}} \cdot R_u^{\text{feedback}} \; + \; w_{\text{ret}} \cdot R_u^{\text{ret}},$$

em que w_{feedback} e w_{ret} são hiperparâmetros que controlam a importância relativa dos dois aspectos (por padrão, $w_{\text{feedback}} = 1.0$ e $w_{\text{ret}} = 0.1$).

Essa formulação permite avaliar o equilíbrio entre exploração e aprofundamento sob uma ótica ampliada: não apenas considerando a resposta imediata ao item recomendado, mas também o impacto indireto na propensão do usuário a retornar em sessões futuras. Além disso, o ambiente registra métricas de **cobertura** e **diversidade intra-lista** (*Intra-List Diversity* – ILD), fornecendo informações adicionais sobre diversidade e dispersão do catálogo.

6.6.6 Comparação de interfaces

A Tabela 12 sumariza as principais diferenças entre os ambientes originais (*Whole-Session* e *Cross-Session*) e o ambiente unificado proposto. Enquanto a versão original separava a modelagem da resposta imediata e da retenção em módulos distintos, o novo ambiente integra ambos em um único ciclo de simulação. Com isso, sinais temporais como δ_t (intervalo entre sessões) e δ_m (média histórica de intervalos) passam a fazer parte do contexto observado pelos agentes, permitindo que algoritmos contextuais explorem diretamente a dinâmica de retorno dos usuários. Além disso, a API foi simplificada para atender *bandits* contextuais, oferecendo uma interface mais direta para observação, ação e cálculo de recompensas.

Tabela 12 – Comparação entre os ambientes originais do KuaiSim e o ambiente unificado proposto.

Aspecto	**Original** (*Whole/Cross*)	**Unificado** (*TimeAwareEnvironment*)
Contexto observado	Não considera sinais temporais entre sessões	Inclui explicitamente δ_t e δ_m no contexto
Saída e retorno	Saída de sessão e retorno modelados de forma separada	Saída acoplada à amostragem de δ_t, garantindo continuidade entre sessões
Ação	Seleção de itens a partir da lista de candidatos	Mesmo formato, compatível com estratégias de seleção adaptativa
Recompensa	Baseada apenas na resposta imediata	Combina resposta imediata com efeitos de retenção
Integração geral	Módulos isolados, sem comunicação direta	Módulos integrados em um único ciclo de simulação

6.6.7 Pseudocódigo do ciclo de simulação

O ciclo de simulação do ambiente unificado integra, de forma contínua, três componentes: *(i)* geração de resposta imediata ao *slate* recomendado; *(ii)* decisão de encerramento da sessão; e *(iii)* amostragem do tempo de retorno do usuário. A cada passo, a política escolhe itens, o ambiente produz a reação do usuário, atualiza seu estado e, quando necessário, avança o relógio para a próxima sessão, incorporando δ_t ao contexto subsequente. O Algoritmo 4 resume essa lógica em alto nível.

Algoritmo 4 TimeAwareEnvironment — ciclo de interação (visão de alto nível)

Entrada: conjunto recomendado de itens (slate)
Resposta imediata: estimar a probabilidade de engajamento por item e amostrar os feedbacks do usuário
Atualização intra-sessão: atualizar o estado de engajamento do usuário e registrar a interação no histórico
Verificar encerramento: decidir se a sessão atual terminou para o usuário
if *todos os usuários encerraram a sessão* **then**
    Retenção/retorno: amostrar o intervalo de retorno δ_t de cada usuário com base no estado e nos feedbacks recentes
    Avanço temporal: avançar o relógio da simulação e preparar a próxima sessão
    Estatísticas de tempo: atualizar as estatísticas de retorno (por exemplo, média histórica δ_m)
end if
Contexto para a política: compor o próximo contexto com perfil, histórico e sinais temporais (δ_t, δ_m)
Cálculo de recompensa: computar a recompensa do passo (podendo incluir componente de retenção)
Saída: retornar o contexto atualizado e o feedback do passo para a política

6.7 Síntese

A adoção da simulação online permitiu superar as limitações da avaliação *offline*, oferecendo um ambiente em que recomendações, respostas dos usuários e intervalos de retorno são modelados de forma integrada. O ambiente proposto (*TimeAwareEnvironment*) conecta três aspectos centrais: (i) resposta imediata, (ii) duração da sessão e (iii) tempo de retorno. Com isso, foi possível avaliar políticas exploratórias em condições mais realistas, considerando tanto os efeitos locais de cada recomendação quanto as consequências de longo prazo na cadência de uso.

Resposta à PP2: Como modelar o tempo em CMAB de forma realista?

A principal contribuição deste capítulo foi mostrar que o tempo deve ser tratado como um elemento endógeno da simulação, isto é, o intervalo de retorno δ_t não é apenas observado, mas também simulado a partir de interações reais, seguindo a distribuição empírica de retornos do KuaiRand, que decai rapidamente após poucos dias e se aproxima de uma distribuição geométrica. Além disso, esse retorno é influenciado pela qualidade das recomendações recebidas, de modo que decisões de curto prazo impactam a cadência de uso futuro.

Esse desenho garante que o tempo seja modelado de forma realista, pois reflete tanto o comportamento estatístico observado em plataformas reais quanto a relação causal entre satisfação imediata e retenção. Ao ser incorporado ao contexto das próximas interações, o tempo passa a orientar diretamente o grau de exploração de algoritmos contextuais.

Em síntese, o tempo em CMAB pode ser considerado realista quando:

- ❏ é gerado a partir de distribuições empíricas de retorno observadas em dados reais;
- ❏ responde às interações do usuário, em vez de ser imposto externamente;
- ❏ retorna como informação de contexto, permitindo que políticas ajustem sua exploração de forma adaptativa.

Essa formulação abre caminho para implementação e aplicação do **Time-Aware Lin-Boltzmann**, que utilizará δ_t como sinal explícito para regular sua exploração e é introduzido no capítulo seguinte.

Capítulo 7

Time-Aware LinBoltzmann

Este capítulo tem como objetivo introduzir o algoritmo **Time-Aware LinBoltzmann**, desenvolvido para investigar o papel de informações temporais na modulação da exploração em sistemas de recomendação baseados em *multi-armed bandits* (MAB) com regressão linear. Este capítulo representa o ponto de convergência dos elementos discutidos anteriormente, em especial as limitações observadas na avaliação *offline* e nos métodos de *Off-Policy Evaluation* (Capítulo 5), bem como a adaptação do simulador *KuaiSim* (Capítulo 6), e apresenta de forma detalhada a formulação e avaliação da abordagem proposta.

No contexto dos MAB contextuais, a decisão de qual item recomendar depende do equilíbrio entre duas forças fundamentais: **exploração** e **aprofundamento**. Os algoritmos lineares clássicos, como o **Lin** (puramente guloso), o **LinGreedy** (*epsilon-greedy*), o **LinUCB** (*Upper Confidence Bound*) e o **LinTS** (*Thompson Sampling*), adotam estratégias distintas para balancear esses dois aspectos. Entretanto, tais métodos não consideram diretamente a dimensão temporal do comportamento dos usuários, o que limita sua capacidade de personalizar a intensidade da exploração de acordo com o contexto de uso.

A proposta do **Time-Aware LinBoltzmann** busca preencher essa lacuna. O algoritmo integra informações sobre o intervalo de tempo entre sessões consecutivas de um mesmo usuário (denotado por δ_t) interpretado como um sinal de retenção, capaz de indicar o nível de engajamento do usuário com a plataforma. Esse valor é então utilizado para ajustar dinamicamente o parâmetro de temperatura do mecanismo de exploração por *Boltzmann* (ou *softmax*), de modo a controlar o grau de aleatoriedade na seleção de itens.

De forma intuitiva, quando o usuário retorna após um longo intervalo entre sessões (δ_t elevado), entende-se que há maior incerteza sobre suas preferências atuais, de modo que a

política tende a intensificar a exploração, apresentando itens mais diversos ou ainda pouco observados. Em contrapartida, quando o retorno ocorre em intervalos curtos (δ_t reduzido), o sistema privilegia o aprofundamento, reforçando recomendações alinhadas ao histórico recente do usuário. Dessa maneira, o algoritmo procura adaptar o nível de exploração às dinâmicas temporais de engajamento, promovendo recomendações potencialmente mais relevantes e personalizadas.

Nas seções seguintes, será apresentada a formulação matemática do **Time-Aware LinBoltzmann**, sua integração ao *TimeAwareEnvironment* do simulador *KuaiSim*, a configuração experimental adotada, bem como a análise comparativa de resultados frente às políticas de referência consideradas.

7.1 Formulação do Time-Aware LinBoltzmann

O algoritmo **Time-Aware LinBoltzmann** é uma extensão do método base **Lin** e do controle da exploração a partir da formulação de *Boltzmann*. Nesta seção, apresenta-se a formulação matemática da proposta, destacando como a informação temporal, representada pelo intervalo entre sessões (δ_t), é incorporada ao processo de decisão.

7.1.1 Revisão do Lin e LinBoltzmann

No **Lin** (*linear bandit* puramente guloso), a política de recomendação baseia-se na regressão linear contextual. Em cada passo de decisão t, para cada item $a \in \mathcal{A}$, estima-se a recompensa esperada como:

$$\hat{r}_{t,a} = \mathbf{x}_{t,a}^{\top} \hat{\boldsymbol{\theta}}_t,$$

onde $\mathbf{x}_{t,a} \in \mathbb{R}^d$ representa o vetor de contexto do par usuário–item e $\hat{\boldsymbol{\theta}}_t$ é o vetor de parâmetros estimado até o instante t. A escolha do item segue a regra:

$$a_t = \arg\max_{a \in \mathcal{A}} \hat{r}_{t,a},$$

caracterizando uma política apenas de aprofundamento.

No proposto **LinBoltzmann**, a decisão é probabilística: em vez de selecionar apenas o item de maior valor estimado, utiliza-se a distribuição *softmax*, que introduz exploração de forma controlada pelo parâmetro de temperatura τ:

$$P(a \mid \mathbf{x}_t) = \frac{\exp\left(\frac{\hat{r}_{t,a}}{\tau}\right)}{\sum_{a' \in \mathcal{A}} \exp\left(\frac{\hat{r}_{t,a'}}{\tau}\right)}.$$

Valores menores de τ tornam a distribuição mais concentrada no item de maior recompensa estimada (comportamento guloso), enquanto valores maiores de τ aumentam a aleatoriedade, favorecendo a exploração de itens alternativos.

7.1.2 Incorporação da dimensão temporal

A principal inovação do **Time-Aware LinBoltzmann** consiste em ajustar dinamicamente a temperatura τ de acordo com o intervalo de tempo entre sessões consecutivas de um usuário, denotado por δ_t. Este intervalo corresponde ao número de dias decorridos entre o final da sessão $t-1$ e o início da sessão t. Assim, considera-se que:

- Um δ_t **elevado** indica que o usuário demorou mais tempo para retornar, sugerindo menor engajamento ou mudança potencial em suas preferências. Nessa situação, a política deve privilegiar exploração, testando novos itens.
- Um δ_t **reduzido** indica que o usuário retornou rapidamente, reforçando a hipótese de que seu interesse permanece estável. Nesse caso, é mais adequado privilegiar o aprofundamento, reforçando recomendações já conhecidas e bem avaliadas.

7.1.3 Função de temperatura adaptativa

O diferencial do **Time-Aware LinBoltzmann** está na forma como o parâmetro de temperatura (τ), responsável por regular o equilíbrio entre exploração e aprofundamento dentro da *softmax*, é ajustado dinamicamente a partir de informações temporais. Em vez de ser mantida constante, a temperatura passa a depender do intervalo de tempo entre sessões consecutivas de um mesmo usuário.

A seguir, são descritas as diferentes estratégias propostas para o cálculo desse parâmetro.

Cálculo da temperatura base

A primeira forma investigada para o cálculo da *temperatura base* τ_t^{base} consiste em uma escala linear pelo tempo de retorno:

$$\tau_t^{\text{base}} = \rho \cdot \delta_t,$$

onde ρ é um fator de escala. Nesse caso, quanto maior o intervalo de tempo desde a última sessão do usuário (δ_t), maior será a temperatura atribuída. O efeito prático é incentivar maior exploração em situações em que o usuário esteve ausente por mais tempo, enquanto valores menores de δ_t mantêm a exploração em níveis reduzidos, favorecendo a exploração apenas quando há indícios de maior incerteza sobre suas preferências atuais.

Ajuste pela variância dos braços

Após o cálculo da temperatura base, considera-se a variabilidade dos valores de recompensa estimados para o conjunto de itens de uma determinada sessão. Seja $\hat{r}_{t,a}$ a

recompensa estimada para o item a no instante t, e seja Var_t a variância dessas recompensas ao longo de todos os itens disponíveis na sessão:

$$\mathrm{Var}_t \;=\; \mathrm{Var}\Big(\{\hat{r}_{t,a} : a \in \mathcal{A}\}\Big).$$

A temperatura ajustada é então definida como:

$$\tau_t \;=\; \tau_t^{\text{base}} \cdot \Big(1 + \alpha \cdot \mathrm{Var}_t\Big),$$

onde $\alpha \geq 0$ é um hiperparâmetro que controla o peso dado à dispersão das estimativas.

A inspiração para esse mecanismo vem de trabalhos como (WANG; ZARIPHOPOULOU; ZHOU, 2020), que analisam a exploração em aprendizado por reforço contínuo e mostram que a variância da distribuição de ações pode ser interpretada como um componente central da exploração. Nesse sentido, sessões em que as estimativas de recompensa apresentam maior variabilidade recebem uma temperatura mais elevada, ampliando o grau de aleatoriedade do *softmax* e, consequentemente, favorecendo a exploração de alternativas diversas.

Regularização pela entropia

Após o cálculo das probabilidades de escolha via distribuição *softmax*,

$$P_t(a) \;=\; \frac{\exp\Big(\hat{r}_{t,a}/\tau_t(a)\Big)}{\sum\limits_{a'} \exp\Big(\hat{r}_{t,a'}/\tau_t(a')\Big)},$$

calcula-se a entropia da distribuição:

$$H_t \;=\; -\sum_{a \in \mathcal{A}} P_t(a)\, \log\Big(P_t(a)\Big).$$

Para manter o nível de estocasticidade em uma faixa desejável, a temperatura é reajustada de acordo com limites inferior ($H_{\min}$) e superior ($H_{\max}$) da entropia:

$$\tau_t(a) \leftarrow \begin{cases} \tau_t(a) \cdot 0.9, & \text{se } H_t > H_{\max}, \\ \tau_t(a) \cdot 1.1, & \text{se } H_t < H_{\min}, \\ \tau_t(a), & \text{caso contrário.} \end{cases}$$

Os fatores multiplicativos 0.9 (redução) e 1.1 (aumento) foram definidos de forma fixa, de modo a fornecer apenas um ajuste incremental e estável da temperatura. Esses valores não foram explorados como hiperparâmetros, uma vez que o objetivo era apenas evitar que a entropia se desviasse excessivamente da faixa $[H_{\min}, H_{\max}]$, e não otimizar essa dinâmica de ajuste.

A noção de regularização por entropia também se apoia em (WANG; ZARIPHOPOULOU; ZHOU, 2020), que analisam formalmente a inclusão da entropia como termo de

regularização para capturar o *trade-off* entre exploração e aprofundamento, mostrando que distribuições com maior entropia correspondem a níveis mais altos de exploração. Esse mecanismo atua, portanto, como um regulador da distribuição de probabilidades, evitando que ela se torne excessivamente determinística (baixa entropia) ou aleatória demais (alta entropia).

A função de temperatura adaptativa combina, portanto, três componentes complementares: (i) o ajuste temporal a partir do intervalo de retorno entre sessões, (ii) o ajuste pela variabilidade dos itens e (iii) a regularização pela entropia. Em conjunto, esses elementos permitem que o algoritmo ajuste de forma mais sensível e personalizada o balanço entre exploração e aprofundamento, capturando tanto a dinâmica temporal do usuário quanto a incerteza associada aos itens disponíveis.

7.1.4 Regra de decisão final

A etapa de decisão do **Time-Aware LinBoltzmann** ocorre em duas fases:

1. **Cálculo da temperatura efetiva.** Primeiro, define-se uma temperatura base τ_t^{base} a partir do intervalo de retorno entre sessões (δ_t). Em seguida, essa temperatura é ajustada pela variância global dos escores estimados na sessão, resultando em:

$$\tau_t^{\text{var}} \;=\; \tau_t^{\text{base}} \cdot \Big(1 + \alpha \cdot \text{Var}_t\Big).$$

   Finalmente, aplica-se o mecanismo de regularização pela entropia, que corrige τ_t^{var} de forma multiplicativa sempre que a entropia da distribuição resultante se afasta dos limites $[H_{\min}, H_{\max}]$. O valor obtido após esse processo é denominado *temperatura efetiva* τ_t^*.

2. **Amostragem de itens via *softmax*.** Com base na temperatura efetiva τ_t^*, define-se a distribuição de probabilidade sobre os itens disponíveis:

$$P_t(a) \;=\; \frac{\exp\Big(\frac{\hat{r}_{t,a}}{\tau_t^*}\Big)}{\sum\limits_{a' \in \mathcal{A}} \exp\Big(\frac{\hat{r}_{t,a'}}{\tau_t^*}\Big)}.$$

   O conjunto de itens a ser recomendado (*slate*) é então obtido por amostragem desta distribuição, sem reposição, até o preenchimento do número máximo de posições disponível.

Dessa forma, a política não apenas utiliza estimativas lineares de recompensa, mas também adapta o nível de aleatoriedade de suas escolhas em função do comportamento temporal de cada usuário, da dispersão das estimativas entre os itens e da regularização pela entropia. O resultado é um processo de decisão mais flexível, capaz de ajustar continuamente o balanço de exploração conforme a dinâmica de retorno às sessões.

7.2 Configuração experimental

Nesta seção são descritos os procedimentos adotados para a avaliação do algoritmo proposto e das políticas de referência selecionadas. O objetivo é explicitar as escolhas de configuração que asseguram a reprodutibilidade dos experimentos e a comparabilidade entre os métodos. São apresentadas as políticas utilizadas como *benchmarks*, o ambiente de simulação empregado, os parâmetros de execução e as métricas consideradas para análise. Essa organização permite estabelecer uma base sólida para a interpretação dos resultados discutidos na próxima seção.

7.2.1 *Benchmarks*

Com o intuito de comparar a proposta **Time-Aware LinBoltzmann** com métodos lineares consolidados, foram conduzidos experimentos com quatro políticas de referência: **Lin** (puramente gulosa), **LinGreedy** (estratégia ϵ-gulosa), **LinUCB** (Upper Confidence Bound) e **LinTS** (Thompson Sampling). Todas as execuções foram realizadas no mesmo ambiente de simulação, seguindo protocolos idênticos em termos de número de iterações, tamanho dos *slates* e sementes aleatórias, garantindo assim comparabilidade direta entre os resultados.

Parâmetros gerais

Os experimentos foram realizados seguindo as seguintes configurações:

- Número de iterações: 20.000.
- Parâmetro de regularização: $\lambda = 1.0$.
- Dimensão do vetor de características lineares: 64 (fixo).
- Tamanhos de *slate*: $\{1, 5, 10, 15, 20\}$.
- Semente: 42.

Configuração específica de cada benchmark

- **Lin**: política gulosa, baseada apenas na regressão linear contextual. Não há hiperparâmetros adicionais além de λ.
- **LinGreedy**: política ϵ-gulosa, que com probabilidade ϵ seleciona um item aleatório e, com $1 - \epsilon$, escolhe o item de maior valor estimado. Foi considerado $\epsilon = 0.1$.
- **LinUCB**: política baseada em intervalo de confiança superior. O hiperparâmetro que controla o tamanho da região de confiança foi definido como $\alpha = 0.1$.

- **LinTS**: política baseada em amostragem de Thompson. O parâmetro de variância da distribuição a priori foi definido como $\nu = 0.1$.

7.2.2 Time-Aware LinBoltzmann

Para a avaliação da proposta **Time-Aware LinBoltzmann**, foi adotado um conjunto fixo de hiperparâmetros, definidos da seguinte forma:

- **Fator de escala da temperatura** (ρ): 1.0.
- **Parâmetro de regularização linear** (λ): 1.0.
- **Peso da variabilidade das estimativas** (α): 1.0.
- **Limites da entropia**: $H_{\min} = 2.0$ e $H_{\max} = 6.0$.

Reconhece-se como limitação deste trabalho a ausência de uma exploração sistemática dos hiperparâmetros, tanto nas políticas de referência quanto no **Time-Aware LinBoltzmann**. Embora a calibração possa impactar o desempenho absoluto dos algoritmos, optou-se por manter valores comumente adotados na literatura ou em estudos preliminares. Essa decisão visou reduzir a variabilidade experimental e concentrar a análise no aspecto central da pesquisa: o efeito da adaptação temporal da temperatura.

Além disso, a busca exaustiva de hiperparâmetros em todos os cenários avaliados implicaria em custos computacionais elevados, o que não se mostrou viável dentro do escopo desta dissertação. Assim, os resultados devem ser interpretados como comparações relativas sob um regime padronizado de configuração, e não como desempenhos máximos possíveis de cada política. Como perspectiva futura, recomenda-se a aplicação de técnicas de ajuste de hiperparâmetros, como busca em grade ou otimização *Bayesiana*, de modo a complementar a análise aqui apresentada.

7.3 Resultados

Nesta seção, são apresentados e analisados os resultados obtidos nos experimentos que compararam o algoritmo proposto **Time-Aware LinBoltzmann** com métodos de referência amplamente utilizados na literatura de *bandits* lineares contextuais: **Lin**, **LinGreedy**, **LinUCB** e **LinTS**. O objetivo principal desta análise é verificar em que medida a adaptação temporal da exploração pode produzir ganhos em métricas relacionadas à diversidade e cobertura, sem comprometer de forma significativa a recompensa acumulada.

Para conduzir a comparação, foram utilizadas seis métricas complementares, cada uma capturando diferentes aspectos do desempenho das políticas avaliadas:

- **Recompensa média**: mede a eficácia imediata do algoritmo em recomendar itens de maior relevância, sendo uma métrica central em sistemas de recomendação baseados em CMAB.

- **Cobertura**: avalia a proporção de itens distintos que foram recomendados ao longo das simulações. Um valor elevado indica maior capacidade de exploração do espaço de itens disponíveis.

- **Área sob a curva de cobertura (*Coverage* AUC)**: considera não apenas a cobertura final, mas também a rapidez com que novos itens passam a ser explorados ao longo do tempo.

- **Diversidade intra-lista (ILD)**: mede a dissimilaridade média entre os itens apresentados em um mesmo *slate*, refletindo a variedade percebida pelo usuário em cada recomendação.

- **Comprimento médio de sessão**: indica a quantidade média de interações que compõem uma sessão antes que o usuário decida encerrar sua atividade. Essa métrica é utilizada como um indicador indireto de engajamento.

- **Intervalo médio entre sessões**: corresponde ao tempo médio que os usuários levaram para retornar ao sistema após cada sessão. Essa métrica permite avaliar possíveis impactos das recomendações sobre a retenção dos usuários ao longo do tempo.

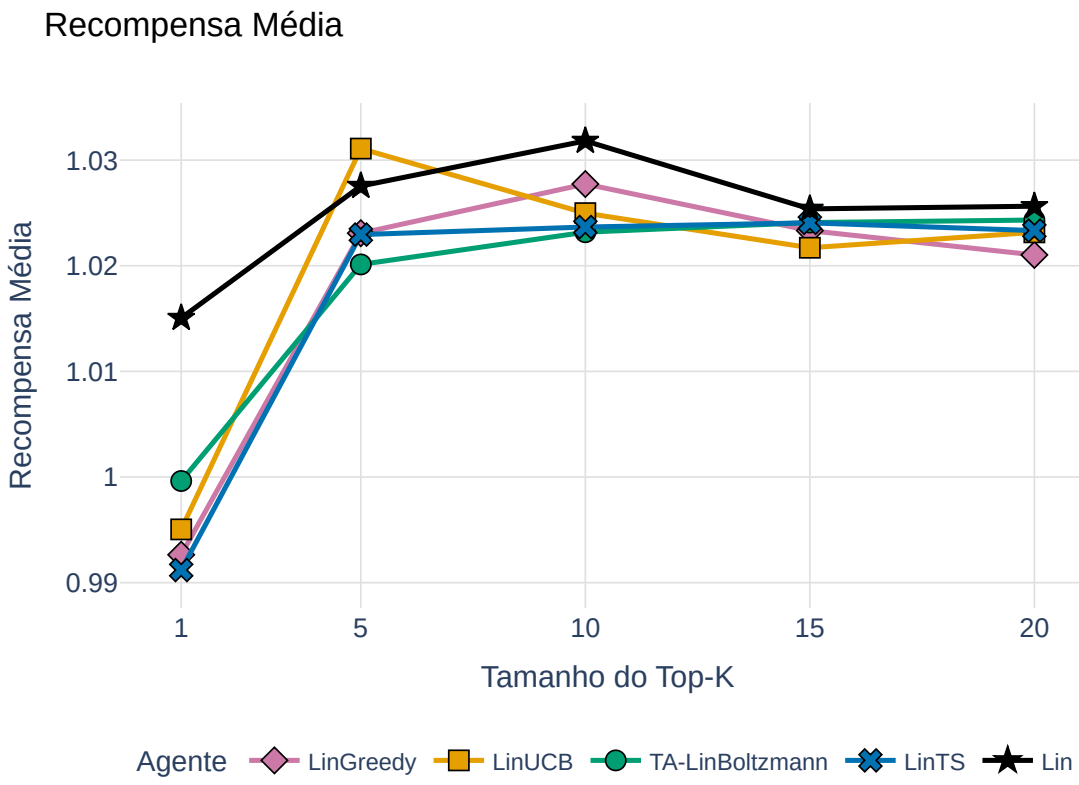
A escolha dessas métricas foi motivada pela necessidade de avaliar não apenas a eficácia em termos de recompensa imediata, mas também aspectos relacionados à exploração do espaço de itens, à experiência percebida pelo usuário e ao comportamento temporal de retorno. Dessa forma, a análise aqui apresentada busca fornecer uma visão abrangente sobre as vantagens e limitações do método temporal em comparação com os métodos de referência.

7.3.1 Recompensa média

A análise inicial considera a recompensa média, métrica que representa o ganho médio obtido a cada interação de recomendação. Esse indicador é fundamental, pois resume a eficácia imediata das políticas em selecionar itens relevantes para os usuários. Os resultados estão apresentados na Figura 7.

Observa-se que a política **Lin**, de natureza puramente gulosa, apresentou o maior valor de recompensa média entre os métodos comparados. No entanto, esse ganho foi relativamente pequeno e bastante próximo ao desempenho das demais políticas que incorporam mecanismos de exploração. Isso sugere que, embora a ausência de exploração permita à

Figura 7 – Recompensa média obtida por cada política.



Fonte: autoria própria.

política concentrar-se nas estimativas de maior retorno imediato, a vantagem obtida em relação às estratégias que equilibram exploração e aprofundamento não é expressiva.

Entre as políticas que realizam algum grau de exploração, incluindo **LinGreedy**, **LinUCB** e **LinTS**, a diferença em relação à política gulosa foi relativamente pequena. O **Time-Aware LinBoltzmann** apresentou valores de recompensa próximos aos *benchmarks* exploratórios, demonstrando que a introdução da adaptação temporal na temperatura de Boltzmann não comprometeu de forma expressiva o desempenho em termos de relevância dos itens recomendados.

Esse resultado é particularmente relevante, pois indica que a proposta deste trabalho (aumentar a exploração de maneira adaptativa ao comportamento temporal dos usuários) não resultou em perdas significativas na métrica central de recompensa. Assim, ainda que não supere o desempenho do método guloso, o método temporal mantém competitividade em eficácia imediata, abrindo espaço para ganhos em outras dimensões, como diversidade e cobertura, analisadas nas próximas subseções.

7.3.2 Cobertura e diversidade

Um segundo conjunto de análises busca avaliar a capacidade das políticas em explorar o espaço de itens disponíveis, indo além da simples maximização da recompensa imediata. Para isso, foram consideradas as métricas de **Cobertura**, **Área Sob a Curva de Cobertura (CoverageAUC)** e **Diversidade Intra-lista (ILD)**. Os resultados podem ser visualizados nas Figuras 8, 9 e 10.

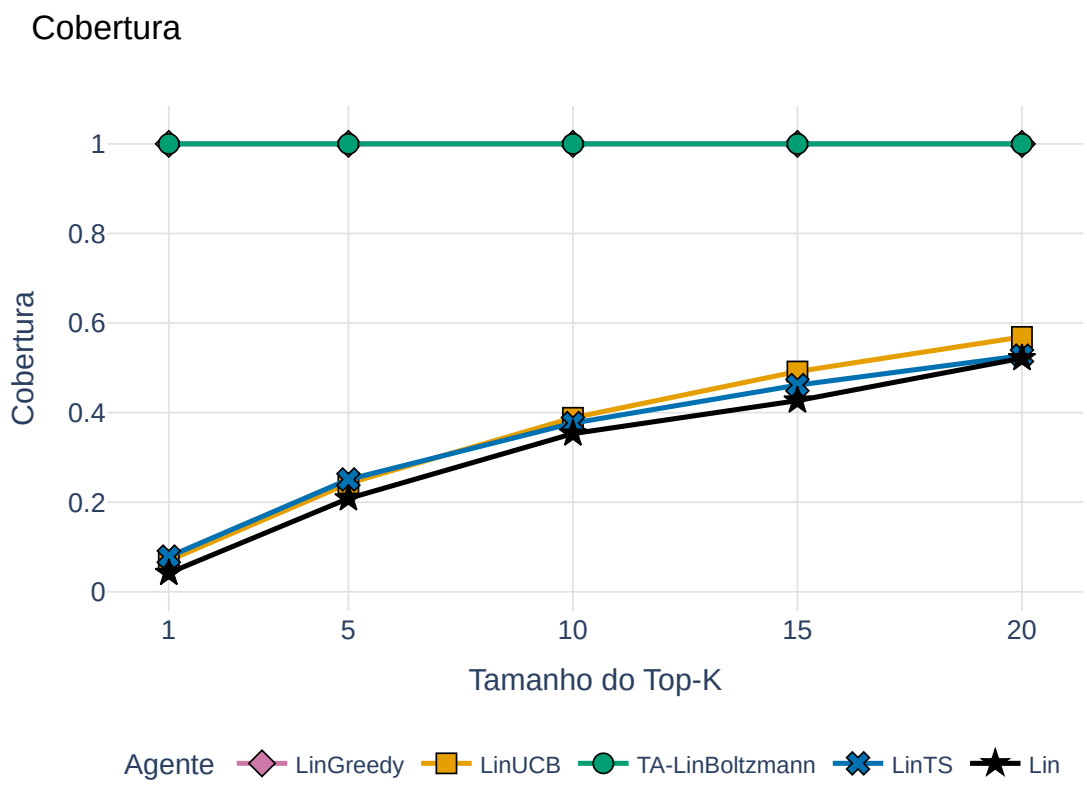
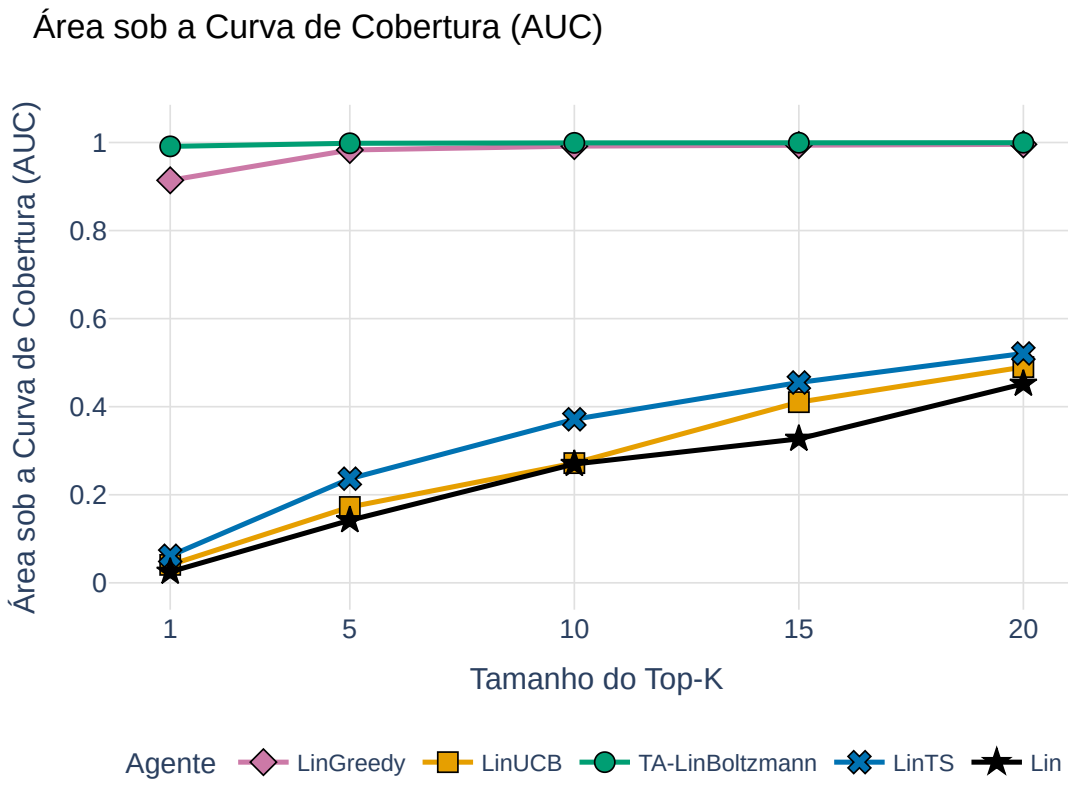


Figura 8 – Cobertura média obtida por cada política.

A Figura 8 mostra que, assim como o **LinGreedy**, o **Time-Aware LinBoltzmann** atingiu cobertura máxima, recomendando a totalidade dos itens do espaço ao longo das simulações. Esse desempenho contrasta fortemente com o obtido pelas demais políticas, que permaneceram restritas a subconjuntos muito menores do catálogo, sobretudo a política **Lin**, cuja cobertura foi bastante limitada em função de seu caráter estritamente guloso.

De forma complementar, a Figura 9 indica que a forma temporal não apenas alcançou maior cobertura final, mas também explorou novos itens de forma mais rápida e consistente ao longo do tempo, obtendo valores de área sob a curva de cobertura próximos ao máximo possível. Esse resultado evidencia a eficiência do mecanismo de exploração adaptativa em distribuir melhor a exposição dos itens desde as etapas iniciais da simulação.

Figura 9 – Área sob a curva de cobertura (CoverageAUC).



Fonte: autoria própria.

Por fim, a Figura 10 apresenta os valores da diversidade intra-lista. Observa-se que o uso do tempo permitiu o método superar todos os *benchmarks* também nesse aspecto, oferecendo *slates* mais variados e com menor redundância entre os itens. Isso sugere que a exploração guiada por sinais temporais favoreceu a geração de recomendações mais diversificadas, o que tende a enriquecer a experiência do usuário.

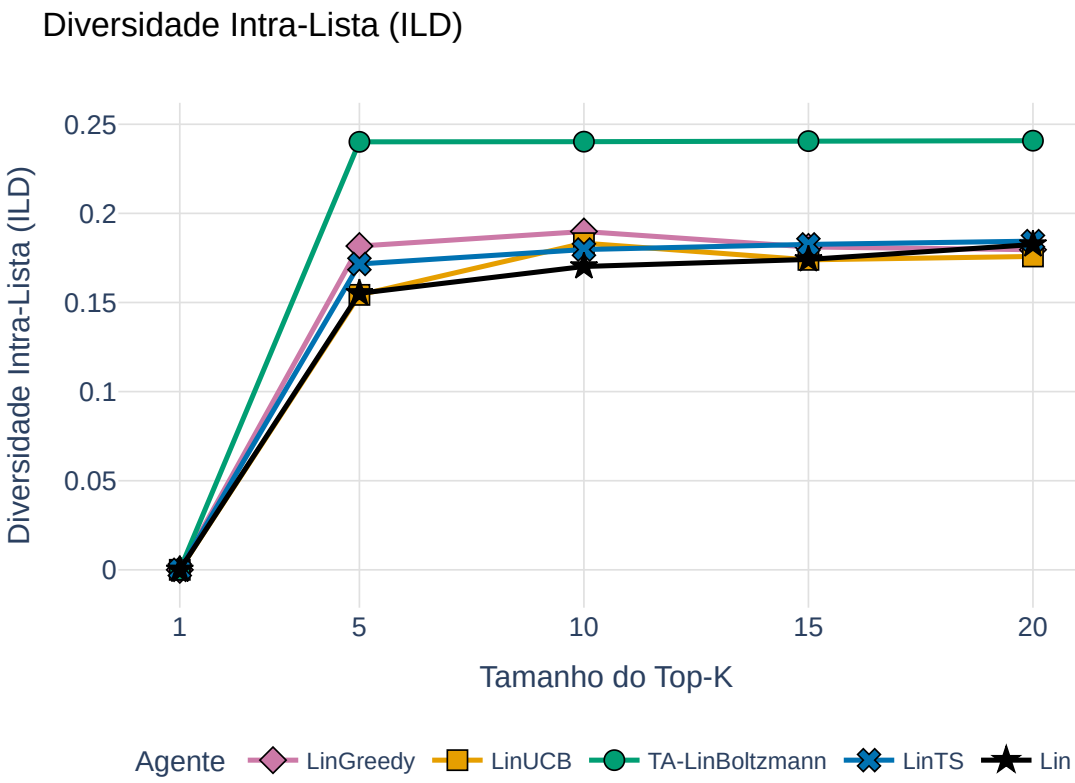
Em síntese, os resultados demonstram que, enquanto as políticas tradicionais alcançaram desempenho limitado em termos de exploração do espaço de itens, o **Time-Aware LinBoltzmann** conseguiu combinar elevada cobertura com diversidade superior, sem prejuízos expressivos em recompensa.

7.3.3 Engajamento e retenção

Além das métricas tradicionais de recompensa, cobertura e diversidade, é importante analisar como as políticas impactam o engajamento e retenção dos usuários simulados. Para isso, foram consideradas duas métricas adicionais: o comprimento médio de sessão, apresentado na Figura 11, e o intervalo médio entre sessões (δ_t), apresentado na Figura 12.

A Figura 11 mostra que o comprimento médio das sessões variou pouco entre os

Figura 10 – Diversidade intra-lista (ILD).



Fonte: autoria própria.

métodos, situando-se em torno de dez interações por sessão. Essa proximidade sugere que as diferentes estratégias de exploração não alteraram de forma significativa a duração média das sessões, indicando um padrão de engajamento relativamente estável.

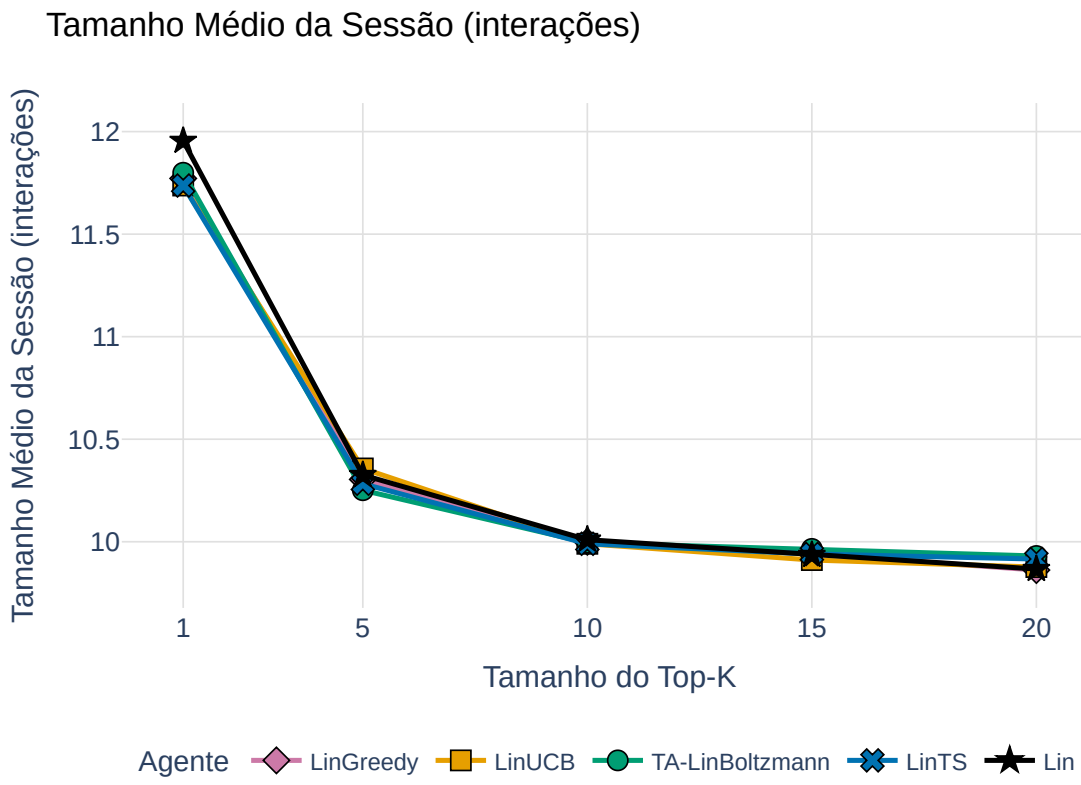
Já a Figura 12 evidencia diferenças quanto ao intervalo médio de retorno dos usuários. O uso do tempo impactou a retenção, apresentando um valor médio maior em comparação aos *benchmarks*. Esse aumento indica que, ao diversificar mais as recomendações, os usuários simulados tenderam a retornar após intervalos um pouco mais longos.

É importante notar que, apesar do método temporal não ter se destacado em relação à retenção, o método puramente guloso também não se destacou, indicando que explorar ainda é positivo. Portanto, o **Time-Aware LinBoltzmann** demonstra ser capaz de equilibrar exploração e aprofundamento sem comprometer de maneira relevante a dinâmica de engajamento dos usuários.

7.3.4 Síntese dos resultados

A análise conjunta das métricas evidencia um panorama claro sobre os pontos fortes e limitações de cada política avaliada. A política **Lin**, por sua natureza estritamente gulosa,

Figura 11 – Comprimento médio de sessão por política.



Fonte: autoria própria.

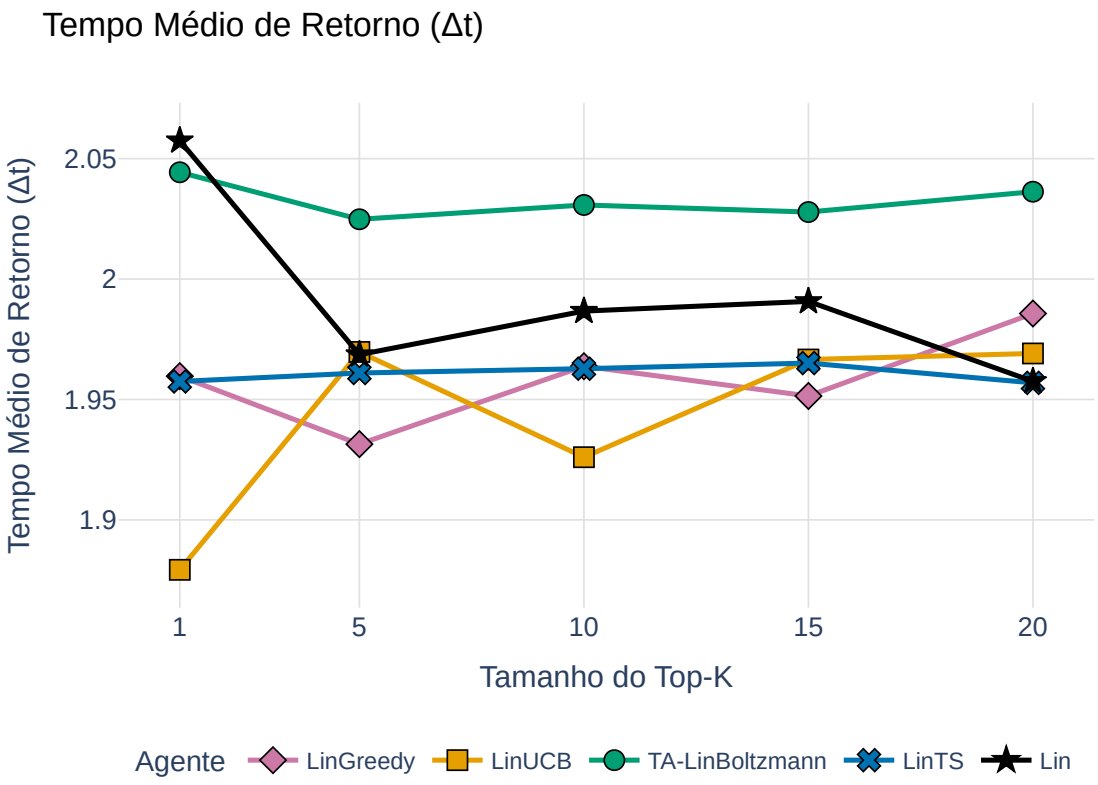
obteve recompensa média um pouco maior, mas apresentou valores bastante reduzidos de cobertura e diversidade. Esse resultado confirma a tendência já discutida de que métodos puramente gulosos favorecem a exploração mínima do espaço de itens, ainda que maximizem a eficácia imediata.

As políticas **LinGreedy**, **LinUCB** e **LinTS**, por sua vez, introduziram algum grau de exploração, alcançando desempenho intermediário em termos de cobertura e diversidade, porém sem se destacarem em nenhuma das dimensões analisadas. Esses métodos servem, portanto, como pontos de referência para comparar a efetividade de estratégias mais sofisticadas de balanceamento entre exploração e aprofundamento.

O **Time-Aware LinBoltzmann** destacou-se de forma consistente nas métricas relacionadas à exploração. Ele atingiu cobertura máxima, apresentou *CoverageAUC* praticamente ideal e obteve os maiores valores de diversidade intra-lista. Esses resultados demonstram que a adaptação da temperatura de Boltzmann com base no intervalo temporal entre sessões favoreceu a exposição de um conjunto mais amplo e variado de itens, ampliando o potencial de descoberta dentro do sistema.

Ainda que tenha apresentado um pequeno aumento no intervalo médio de retorno dos usuários, esse efeito foi modesto e não comprometeu de maneira significativa o en-

Figura 12 – Intervalo médio entre sessões (δ_t) por política.



Fonte: autoria própria.

gajamento, como sugerido pelo comprimento médio de sessão, que permaneceu estável em relação aos *benchmarks*. Em termos de recompensa, o *Time-Aware LinBoltzmann* manteve-se competitivo, com valores próximos às políticas exploratórias, confirmando que os ganhos em diversidade e cobertura foram alcançados sem perdas expressivas de eficácia imediata.

Em síntese, os resultados corroboram a hipótese central desta pesquisa: informações temporais podem ser utilizadas de maneira eficaz para modular a exploração em algoritmos de *bandits* contextuais. Dessa forma, em resposta à **PP3 – Qual o efeito do tempo como fator de exploração?**, conclui-se que o uso do tempo aumentou de forma expressiva a diversidade e a cobertura, ao mesmo tempo em que manteve a recompensa em níveis competitivos. O **Time-Aware LinBoltzmann** apresentou-se, portanto, como a alternativa mais equilibrada entre todas as políticas avaliadas, conciliando relevância, diversidade e cobertura em um mesmo modelo.

7.4 Alternativas para o uso do tempo

O cálculo da temperatura desempenha papel central no mecanismo de exploração do *Time-Aware LinBoltzmann*. Os resultados apresentados até então consideram apenas uma forma linear de ajuste baseada no intervalo entre sessões (δ_t), em que valores maiores de tempo de retorno levam a temperaturas mais altas e, portanto, a níveis mais intensos de exploração. Essa formulação simples e direta mostrou-se adequada como ponto de partida.

Com o objetivo de explorar possibilidades adicionais, foram investigadas três variantes no uso do tempo. Essas propostas não constituem contribuições centrais desta dissertação, mas sim uma análise preliminar e exploratória de como o fator temporal poderia ser incorporado de maneira alternativa. Entre as motivações estão a possibilidade de ajustar o impacto relativo de δ_t de acordo com o histórico de cada usuário e a suavização de efeitos extremos, como retornos excepcionalmente longos que levam a temperaturas desproporcionais.

Dessa forma, tais variantes devem ser entendidas como sugestões iniciais para modificações no cálculo da temperatura, cujo interesse é ampliar a discussão sobre abordagens temporais, sem pretensão de oferecer resultados conclusivos ou superiores ao modelo linear de referência.

Formas de ajustar a temperatura

Foram investigadas três alternativas para o cálculo da *temperatura base* τ_t^{base}. Cada uma delas procura explorar o intervalo de retorno entre sessões (δ_t) de maneira distinta, oferecendo diferentes implicações para o equilíbrio entre exploração e aprofundamento.

1. **Escala linear pelo tempo de retorno**:

$$\tau_t^{\text{base}} = \rho \cdot \delta_t,$$

   método já apresentado, onde ρ é um fator de escala. Nesse caso, quanto maior o intervalo desde a última sessão, maior será a temperatura, incentivando diretamente a exploração.

2. **Escala relativa ao histórico do usuário**:

$$\tau_t^{\text{base}} = \rho \cdot \frac{\delta_t}{\delta_m},$$

   em que δ_m representa o intervalo médio de retorno do próprio usuário. Essa formulação torna o ajuste personalizado: o mesmo valor absoluto de δ_t pode gerar temperaturas diferentes de acordo com o comportamento histórico de cada indivíduo.

3. **Escala logarítmica**:

$$\tau_t^{\text{base}} = \rho \cdot \log\left(1 + \frac{\delta_t}{\delta_m}\right).$$

Nessa versão, o crescimento da temperatura é suavizado para grandes valores de δ_t, evitando que retornos muito longos provoquem níveis de exploração desproporcionais em relação ao restante do histórico.

Cada uma dessas variantes apresenta potenciais benefícios em contextos específicos. Enquanto a forma linear é mais direta e favorece fortemente a exploração após longos períodos sem interação, a forma relativa permite adaptações personalizadas por usuário, e a logarítmica atua como um mecanismo de suavização para cenários com grande variabilidade de tempo entre sessões.

Protocolo de experimentação e avaliação

A avaliação das diferentes variantes de temperatura foi conduzida por meio de experimentos sistemáticos, garantindo a cobertura de uma ampla combinação de parâmetros. O objetivo principal foi explorar o espaço de configurações possíveis, observando como cada método de cálculo da temperatura se comporta em diferentes cenários do ambiente simulado.

Execução dos experimentos

Cada execução do *Time-Aware LinBoltzmann* foi realizada com $n_{\text{iter}} = 20{,}000$ interações e com semente fixa (42), de modo a assegurar a reprodutibilidade dos resultados. Os parâmetros variaram de acordo com os seguintes conjuntos:

- ❏ **Tamanho do *slate*** (k): $\{1, 5, 10, 15, 20\}$;
- ❏ **Fator de escala da temperatura** (ρ): $\{0.5, 1.0, 2.0\}$;
- ❏ **Método de cálculo da temperatura**: $\{\delta_t, \delta_t/\delta_m, \log(1 + \delta_t/\delta_m)\}$;
- ❏ **Peso da variabilidade dos escores** (α): $\{0.1, 1.0\}$;
- ❏ **Limites de entropia**: $H_{\min} \in \{2.0, 3.0\}$ e $H_{\max} \in \{5.0, 6.0\}$.

O parâmetro de regularização λ foi mantido fixo em 1,0, e o mecanismo de ajuste por variância e entropia permaneceu sempre ativado. A combinação completa desses valores resultou em um total de 360 execuções independentes.

Protocolo de avaliação dos resultados

Após a execução dos experimentos, cada configuração gerou trajetórias com as interações simuladas entre usuários e recomendações. Para comparar de forma justa as diferentes variantes de cálculo da temperatura, foi definido um procedimento de avaliação em duas etapas: (i) cálculo e normalização das métricas em cada cenário e (ii) seleção do conjunto de hiperparâmetros mais consistente para cada método de temperatura.

As métricas utilizadas foram as mesmas apresentadas anteriormente: recompensa média, cobertura, área sob a curva de cobertura, diversidade intra-lista, comprimento médio da sessão e tempo médio de retorno. Como o valor de k influencia diretamente esses resultados, aplicou-se uma normalização do tipo *min–max* separada para cada valor de k. Para as métricas em que valores maiores indicam melhor desempenho, a normalização foi feita diretamente; já no caso do tempo médio de retorno, cujo menor valor é preferível, a normalização foi aplicada sobre o valor invertido ($-\delta_t$).

Em seguida, foi calculada uma pontuação média para cada configuração e para cada k, denominada Pontuação_k, obtida pela média aritmética das métricas normalizadas. Assim, todas as dimensões avaliadas receberam o mesmo peso, refletindo um balanço entre desempenho, diversidade, cobertura e aspectos temporais.

Para cada método de temperatura e cada valor de k, as combinações de hiperparâmetros foram ordenadas segundo a Pontuação_k. Em casos de empate, os critérios de desempate seguiram a ordem: recompensa média, área sob a curva de cobertura, cobertura, diversidade intra-lista, comprimento médio da sessão (em ordem decrescente) e tempo médio de retorno (em ordem crescente).

A partir dessa ordenação, foi selecionada, para cada método de temperatura, a configuração de hiperparâmetros mais consistente ao longo de todos os valores de k. A consistência foi avaliada pelo número de vezes em que a configuração ocupou a primeira posição, pela média e pior posição registrada, além da pontuação média $\overline{\text{Pontuação}}_k$ e sua variabilidade.

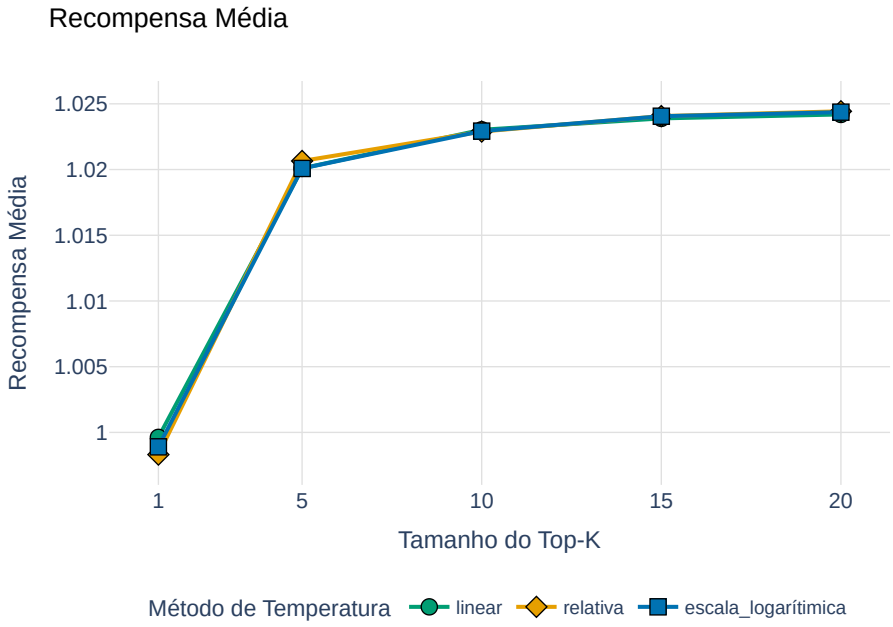
Esse procedimento garantiu que a escolha dos hiperparâmetros não dependesse de um único valor de k, mas sim do desempenho global da configuração em diferentes cenários. Assim, foi possível identificar para cada método de cálculo da temperatura um conjunto único de parâmetros representativo e robusto para a análise comparativa posterior.

Métricas de desempenho e diversidade

A Figura 13 mostra os valores de recompensa média obtidos por cada variante de temperatura. Nota-se que as curvas seguem praticamente o mesmo padrão ao longo dos diferentes tamanhos de *slate*, com diferenças pouco expressivas entre os métodos. De forma análoga, os resultados de cobertura (Figura 14), área sob a curva de cobertura (Figura 15), e diversidade intra-lista (ILD) (Figura 16), revelam desempenhos equivalentes,

com valores praticamente sobrepostos. O mesmo ocorre com o comprimento médio de sessão (Figura 17), cujas variações entre os métodos são marginais.

Figura 13 – Recompensa média (*AverageReward*) para os diferentes métodos de cálculo da temperatura.



Fonte: autoria própria.

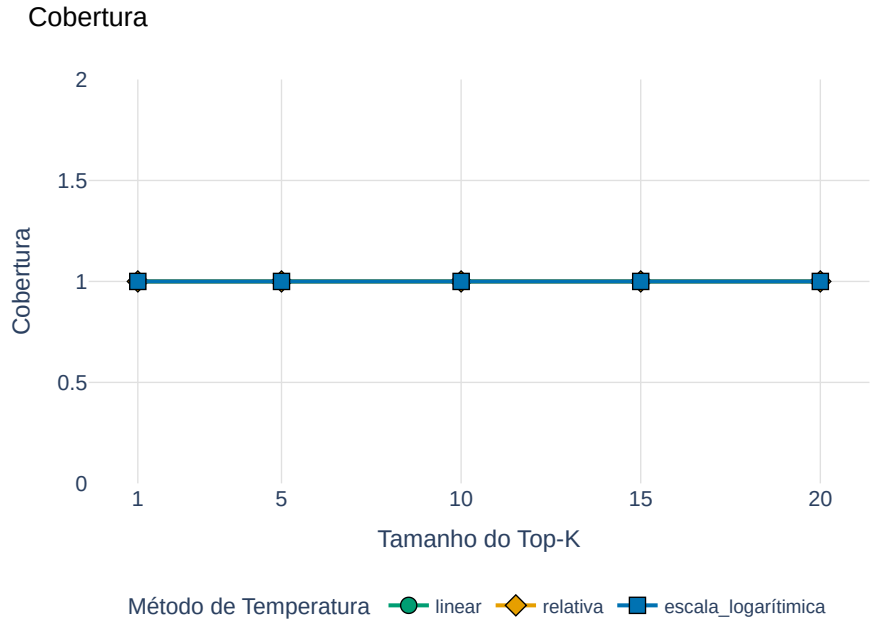
Métrica temporal

Entre todas as métricas analisadas, a que apresentou variação mais perceptível foi o tempo médio de retorno, conforme ilustrado na Figura 18. Embora as diferenças entre os métodos não sejam numericamente expressivas, situando-se na ordem de centésimos de unidade, elas se mantiveram consistentes ao longo dos diferentes cenários de avaliação. Esse resultado corrobora a conclusão já destacada anteriormente, quando da análise dos *benchmarks* e do **Time-Aware LinBoltzmann** linear: o tipo de exploração parece exercer um certo impacto sobre a retenção (intervalo de retorno dos usuários).

Síntese dos hiperparâmetros vencedores

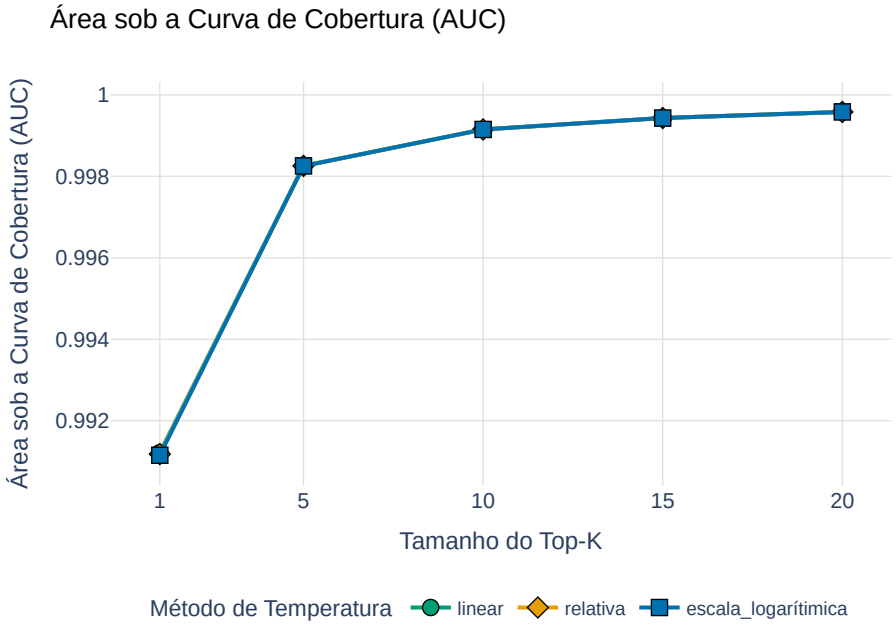
A partir da consolidação dos resultados ao longo de todos os valores de k, foi selecionado, para cada método de temperatura, um único conjunto de hiperparâmetros mais consistente. A Tabela 13 apresenta os parâmetros destacados.

Figura 14 – Cobertura média (*Coverage*) obtida pelos métodos de temperatura.



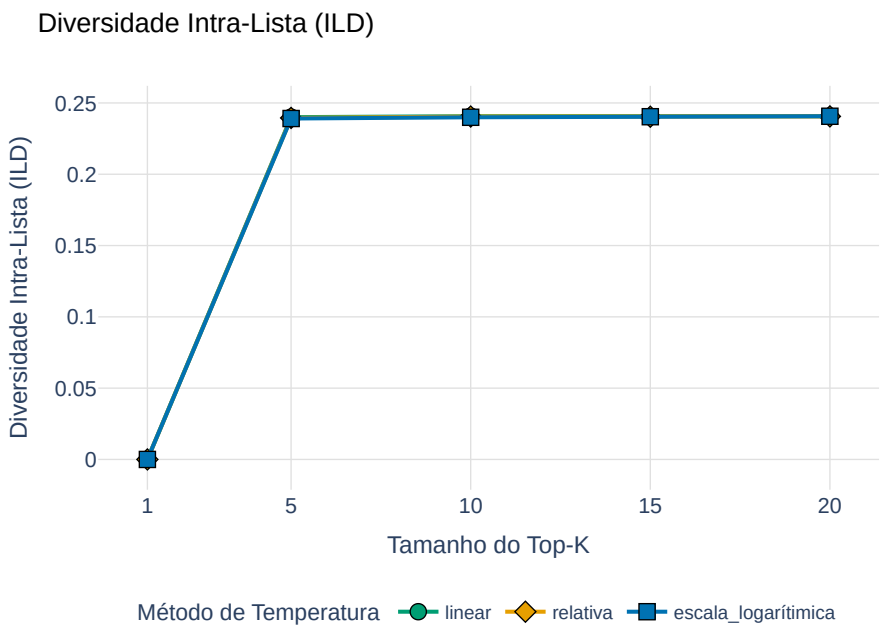
Fonte: autoria própria.

Figura 15 – Área sob a curva de cobertura (*CoverageAUC*).



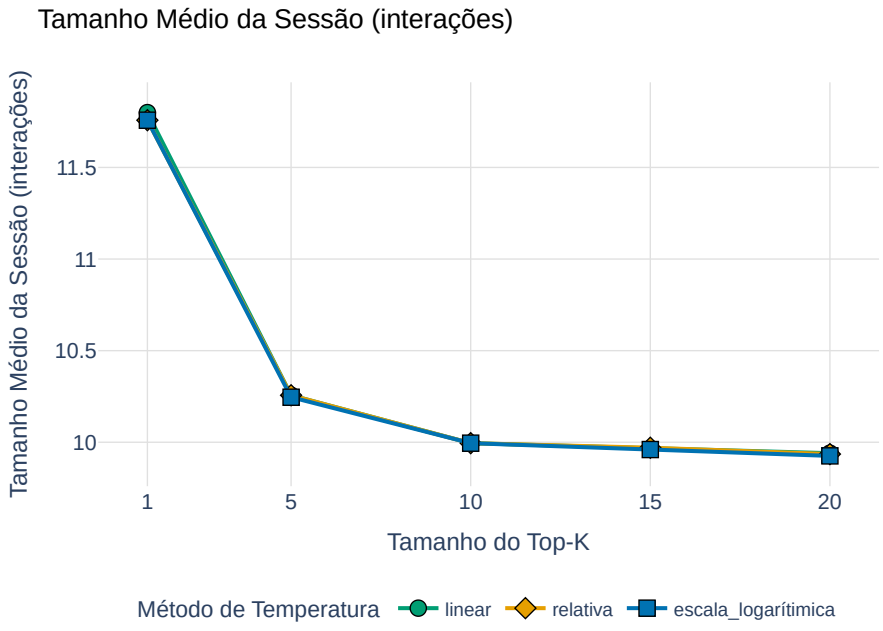
Fonte: autoria própria.

Figura 16 – Diversidade intra-lista (*ILD*) nos diferentes métodos.



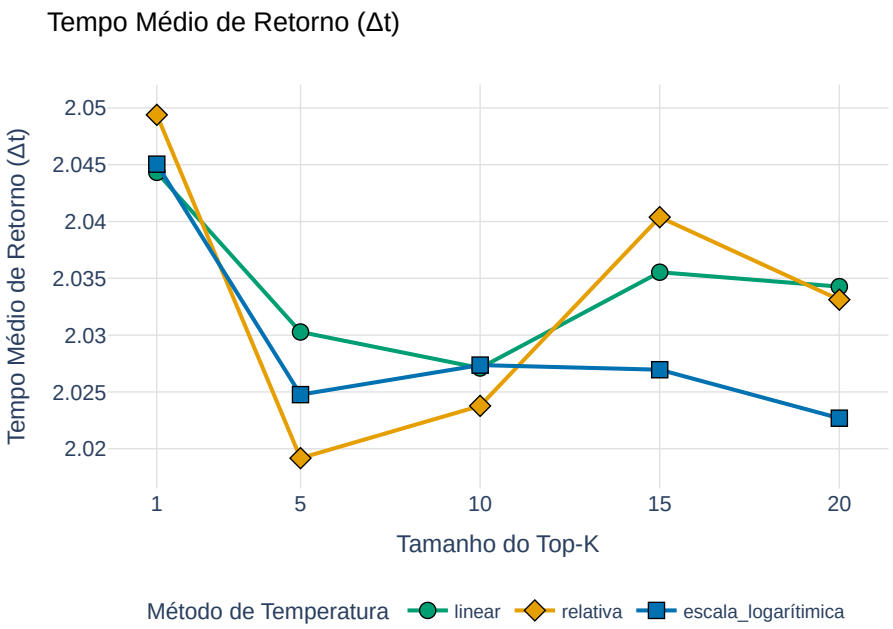
Fonte: autoria própria.

Figura 17 – Comprimento médio de sessão (*AvgSessionLength*) em função do tamanho do *slate*.



Fonte: autoria própria.

Figura 18 – Tempo médio de retorno dos usuários (*AvgDeltaT*), métrica mais sensível à variação do método de temperatura.



Fonte: autoria própria.

Tabela 13 – Conjuntos de hiperparâmetros mais consistentes para cada método de temperatura, considerando todos os valores de k.

Método	ρ	λ	α	$H_{\min}$	$H_{\max}$
δ_t	1,0	1,0	1,0	2,0	5,0
$\log(1+\delta_t/\delta_m)$	1,0	1,0	1,0	3,0	6,0
δ_t/δ_m	1,0	1,0	1,0	3,0	5,0

Considerações finais

O conjunto de experimentos apresentados teve como propósito principal explorar possibilidades iniciais de incorporação do fator temporal no cálculo da temperatura do *Boltzmann*. Os resultados mostram que as formas de variação consideradas (linear, relativa e logarítmica) não produziram ganhos expressivos em métricas clássicas como recompensa média, cobertura ou diversidade. Ainda assim, observou-se que a métrica de retorno foi a mais sensivelmente impactada, indicando que o tempo pode exercer influência relevante sobre a dinâmica de retorno dos usuários. Esse indício sugere que ajustes mais intensos ou não lineares da temperatura, possivelmente baseados em funções mais sofisticadas de δ_t ou em estratégias híbridas, podem revelar efeitos mais pronunciados.

Embora os ganhos imediatos não tenham sido significativos, os resultados obtidos apontam para a necessidade de experimentos mais profundos e sistemáticos a fim de

compreender melhor o real impacto de diferentes formas de exploração sobre a retenção.

Capítulo 8

Discussão

A presente dissertação teve como objetivo central investigar o papel de informações temporais, em especial o intervalo entre interações consecutivas (δ_t), na regulação do equilíbrio entre exploração e aprofundamento em sistemas de recomendação baseados em *Multi-Armed Bandits* contextuais. Partindo da hipótese de que o tempo pode atuar como um modulador dinâmico da exploração, buscou-se avaliar de que forma essa dimensão influencia métricas tradicionais de desempenho, bem como medidas complementares de diversidade e cobertura.

As perguntas de pesquisa formuladas no Capítulo 4 nortearam o desenho experimental e estruturam a presente discussão. De forma resumida, as questões centrais podem ser agrupadas em quatro eixos:

- **Avaliação**: quais são as formas mais adequadas de avaliar algoritmos de MAB contextuais que incorporam exploração adaptativa?
- **Modelagem temporal**: como representar o tempo de retorno dos usuários de modo realista, levando em conta sua dependência do histórico de consumo e das próprias recomendações recebidas?
- **Efeitos práticos**: quais impactos decorrem do uso do tempo como fator de exploração em métricas como recompensa média, diversidade e cobertura?
- **Limitações e desafios**: quais barreiras de dados, protocolos e generalização dificultam a incorporação efetiva do tempo em algoritmos de recomendação baseados em MAB?

A análise crítica dos resultados experimentais obtidos ao longo desta pesquisa permite confrontar essas perguntas com as hipóteses formuladas no Capítulo 4. Assim, a seção

seguinte discute, em primeiro lugar, a questão da avaliação de políticas exploratórias e as limitações intrínsecas dos protocolos *offline*, avançando depois para a modelagem temporal no simulador, os efeitos práticos do uso de informações de tempo e, por fim, os desafios e implicações futuras desse campo de estudo.

8.1 Avaliação *offline* e suas limitações

A primeira pergunta de pesquisa (**PP1**: *Como avaliar de maneira justa CMAB com exploração adaptativa? Quais condições e protocolos são mais adequados*) buscava investigar a adequação da avaliação *offline* para políticas de *Multi-Armed Bandits* contextuais que dependem de exploração. Em particular, a questão central era compreender se protocolos baseados em *logs* históricos estáticos e avaliações contrafactuais seriam capazes de refletir de forma justa o desempenho de políticas que modulam sua exploração ao longo do tempo. Para examinar essa hipótese, foi conduzido um conjunto de experimentos comparando políticas lineares bem estabelecidas (**Lin**, **LinGreedy**, **LinUCB** e **LinTS**) em cenários de avaliação *offline*, sem ainda incluir variações com consciência temporal. O foco inicial, portanto, esteve em verificar a validade metodológica do protocolo, antes de aplicar técnicas de exploração adaptativa baseadas em informação temporal.

Os resultados, detalhados no Capítulo 5, indicaram de forma consistente que a política gulosa (**Lin**), sem qualquer componente exploratório, superou ou igualou o desempenho das demais variantes em praticamente todos os cenários avaliados. Este achado mostra que protocolos *offline* tendem a favorecer o *exploitation*, penalizando estratégias que buscam explorar novos itens. Mesmo com o uso de técnicas de *off-policy evaluation* (OPE), como IPW, DM e DR, a vantagem do modelo guloso se manteve, evidenciando que o viés contra exploração não se limita apenas ao método de puramente offline.

Esse fenômeno foi analisado em profundidade e resultou no artigo *"Exploitation Over Exploration: Unmasking the Bias in Linear Bandit Recommender Offline Evaluation"*, aceito na **19th ACM Conference on Recommender Systems** (*RecSys 2025*). O estudo demonstrou, em larga escala, que em mais de 89% dos casos avaliados em diferentes bases de dados, o modelo Lin obteve desempenho superior ou equivalente às demais políticas exploratórias, confirmando a existência de um viés estrutural nos protocolos de avaliação *offline* de bandits lineares (PIRES et al., 2025).

Essas evidências permitem responder à **PP1**: avaliações *offline* não são adequadas para medir o impacto real de políticas exploratórias em MAB contextuais. Elas induzem uma preferência artificial pelo aprofundamento, desvalorizando os potenciais benefícios da exploração. Esse resultado está em linha com a hipótese formulada (**H1**), que afirmava que protocolos *offline*/OPE tendem a subestimar algoritmos com exploração adaptativa quando baseados em *logs* determinísticos e com suporte limitado. Assim, confirma-se que a avaliação justa desses algoritmos exige metodologias alternativas, como ambien-

tes de simulação interativa, capazes de capturar a influência das recomendações sobre o comportamento futuro dos usuários.

8.2 Avaliação realista do tempo e adaptação do simulador

A segunda pergunta de pesquisa (**PP2**: *Como modelar o tempo em CMAB de forma realista? Em especial, como representar o tempo de retorno quando ele também é influenciado pelas próprias recomendações?*) buscava compreender de que maneira o intervalo entre sessões consecutivas (δ_t) poderia ser tratado como parte integrante do processo de simulação. O objetivo não era apenas registrar o tempo de retorno como um dado auxiliar, mas sim incorporá-lo como uma variável endógena, isto é, condicionada tanto ao histórico de consumo do usuário quanto à qualidade das recomendações recebidas. Essa formulação permitiu avaliar não só a resposta imediata às recomendações, mas também a influência delas sobre a retenção e o ritmo de engajamento futuro.

A solução encontrada foi recorrer ao simulador *KuaiSim*, construído a partir do conjunto de dados *KuaiRand*, o qual contém *timestamps* reais de interações. O diferencial do KuaiSim é que ele inclui um módulo de retenção responsável por estimar o tempo de retorno dos usuários após cada sessão. Esse módulo implementa um modelo probabilístico que, ao término da sessão, amostra um valor de δ_t a partir de uma distribuição geométrica, cuja parametrização depende (i) de um viés global, (ii) de fatores individuais de propensão ao retorno e (iii) da qualidade das recomendações observada na sessão. Essa abordagem reflete evidências empíricas, já que a distribuição de tempos de retorno no *KuaiRand* apresenta decaimento geométrico após poucos dias.

Para os propósitos desta dissertação, o KuaiSim foi profundamente modificado, originando o ambiente *TimeAwareEnvironment*. Essa versão unificou os módulos de resposta imediata e de retenção, permitindo que δ_t não fosse apenas registrado, mas incorporado diretamente ao contexto observado pelos agentes. Assim, políticas contextuais, como o **Time-Aware LinBoltzmann**, puderam utilizar δ_t como sinal explícito para modular sua exploração.

O KuaiSim representa uma opção adequada por utilizar dados reais com modelagem probabilística realista do retorno dos usuários. Entre os simuladores revisados, ele foi o único identificado que integra de forma explícita o tempo entre sessões como parte do ciclo de simulação. Essa característica o torna um recurso valioso para experimentos em que se deseja estudar os efeitos de longo prazo das recomendações e sua influência na cadência de interação.

Dessa forma, confirmam-se os pressupostos da hipótese **H2**, segundo a qual a avaliação do impacto temporal só é realista quando o intervalo de retorno (δ_t) é modelado como variável endógena, dependente tanto do histórico do usuário quanto das recomendações

oferecidas. A adaptação do KuaiSim viabilizou essa abordagem, permitindo conduzir experimentos que, diferentemente dos protocolos *offline*, capturam a relação dinâmica entre exploração, recompensas e comportamento temporal dos usuários.

8.3 Impactos do tempo como fator de exploração

A terceira pergunta de pesquisa (**PP3**: *Qual o efeito do tempo como fator de exploração? Impactos de usar informações temporais na exploração sobre, por exemplo, recompensa, diversidade e cobertura*) tratava de investigar como a variável temporal poderia influenciar diretamente o equilíbrio entre exploração e aprofundamento em algoritmos de MAB contextuais. A hipótese correspondente (**H3**) previa que a incorporação de informações temporais teria potencial para melhorar o desempenho dos algoritmos, ampliando a satisfação do usuário não apenas pela relevância imediata das recomendações, mas também pela sua diversidade e pela capacidade de alcançar uma cobertura mais ampla do espaço de itens.

A análise dos resultados mostrou que, em termos de recompensa média, o **Time-Aware LinBoltzmann** não superou a política puramente gulosa (**Lin**), que se manteve como a mais eficaz no curto prazo (apesar de não ser uma grande diferença). Esse desfecho mostra que a ausência de exploração favorece ganhos imediatos, ainda que à custa de uma experiência menos diversificada. O aspecto positivo, contudo, é que o método temporal apresentou desempenho semelhante aos benchmarks exploratórios (LinGreedy, LinUCB, LinTS), evidenciando que a adaptação temporal não implicou em perdas de recompensa imediata.

Por outro lado, nas métricas de cobertura e diversidade, o uso do tempo trouxe melhoras consistentes. Enquanto as demais políticas permaneceram restritas a subconjuntos relativamente pequenos do catálogo, o método proposto conseguiu explorar praticamente todo o espaço de itens. Esse efeito foi acompanhado de maior diversidade entre os elementos recomendados, indicando que a exploração guiada por sinais temporais levou a uma distribuição mais equilibrada da exposição de itens. Trata-se de um resultado particularmente relevante, pois métricas de diversidade e cobertura estão associadas à descoberta de novos conteúdos e à mitigação de efeitos de concentração excessiva em itens populares, problemas recorrentes em sistemas de recomendação.

No que diz respeito ao comportamento temporal dos usuários simulados, o impacto foi mais sutil. O comprimento médio das sessões mostrou-se estável entre os métodos, sugerindo que a estratégia exploratória não alterou a duração das interações. Já o intervalo médio de tempo entre sessões foi ligeiramente maior utilizando a exploração com tempo, indicando que a diversificação pode ter levado os usuários a retornarem com menor frequência. No entanto, essa diferença foi pequena e não comprometeu os ganhos obtidos em cobertura e diversidade.

De forma geral, os resultados oferecem uma confirmação parcial da hipótese **H3**. Embora a incorporação de informações temporais não tenha produzido ganhos consistentes em termos de recompensa média, observou-se um impacto positivo em métricas ligadas à experiência de uso em longo prazo, como diversidade e cobertura. Esses achados indicam que o tempo pode atuar como um modulador eficaz da exploração, contribuindo para equilibrar relevância imediata e variedade nas recomendações. O desafio que permanece é ampliar tais benefícios sem comprometer o engajamento, de modo que os ganhos em diversidade e cobertura também se traduzam em melhorias perceptíveis em recompensa e retenção.

8.4 Limitações, desafios e potenciais do tempo como fator de exploração

A quarta pergunta de pesquisa (**PP4**: *Quais limitações e desafios práticos dessa abordagem? Restrições de dados, protocolos e generalização*) abordava as restrições e dificuldades de se incorporar o tempo como fator de exploração em sistemas de recomendação baseados em MAB contextuais. A hipótese associada (**H4**) reconhecia que essa incorporação esbarraria em barreiras práticas, como a escassez de dados adequadamente anotados, a ausência de protocolos de avaliação capazes de capturar efeitos de longo prazo e a necessidade de ferramentas específicas, mas defendia que, apesar dessas limitações, o tempo poderia abrir espaço para métodos mais adaptativos e personalizados.

A revisão da literatura já apontava uma lacuna significativa: embora o tempo apareça com frequência como atributo contextual ou como variável auxiliar em técnicas de filtragem colaborativa, são raros os trabalhos de CMAB que o incorporam explicitamente como modulador da exploração. Essa ausência decorre, em grande parte, da dificuldade em obter *logs* com intervalos de retorno representativos e da inexistência de protocolos de avaliação que consigam refletir a dinâmica entre satisfação imediata e engajamento futuro. O presente trabalho confirmou esse diagnóstico ao mostrar que protocolos *offline* penalizam a exploração e, portanto, inviabilizam qualquer análise realista sobre a influência temporal.

Os experimentos realizados também evidenciaram desafios metodológicos. Foi necessário recorrer ao KuaiSim, extensivamente adaptado, para possibilitar a avaliação de políticas *time-aware*. Apesar do realismo introduzido pela modelagem endógena do retorno, o simulador ainda representa um ambiente artificial, cuja fidelidade depende da qualidade do modelo de retenção e da plausibilidade de suas suposições. Assim, embora a simulação tenha viabilizado avanços, permanece a limitação de não se tratar de dados coletados em sistemas reais.

Em relação aos resultados, os ganhos do **Time-Aware LinBoltzmann** não se refletiram na métrica mais tradicional de recompensa média, mas se manifestaram em dimensões

secundárias, como cobertura e diversidade. Esse achado reforça a hipótese **H4**: mesmo diante de restrições de protocolo e de dados, o tempo pode ser explorado como recurso estratégico para enriquecer a experiência do usuário, ampliando a variedade e a capacidade de descoberta sem depender exclusivamente de ganhos imediatos em eficácia.

A principal contribuição deste trabalho, portanto, não está em apresentar um método que supera de forma inequívoca os benchmarks existentes, mas em abrir um caminho para o desenvolvimento de políticas sensíveis ao tempo e para a criação de metodologias de avaliação mais adequadas. Ao evidenciar as limitações das avaliações *offline*, propor modificações no KuaiSim e demonstrar empiricamente os efeitos do tempo sobre métricas de diversidade e cobertura, a pesquisa estabelece um terreno fértil para investigações futuras.

Em síntese, as limitações identificadas (escassez de dados temporais, dependência de simuladores e ausência de protocolos padronizados) não invalidam o uso do tempo como fator de exploração. Pelo contrário, ressaltam a necessidade de novas metodologias e ferramentas que tornem possível construir sistemas capazes de ajustar dinamicamente sua estratégia exploratória ao ritmo de interação de cada usuário, resultando em recomendações mais dinâmicas, personalizadas e alinhadas à cadência real de consumo.

Capítulo 9

Conclusão

Esta dissertação teve como objetivo investigar o impacto da incorporação de informações temporais no balanceamento entre exploração e aprofundamento em sistemas de recomendação baseados em *Multi-Armed Bandits* (MAB) contextuais. A proposta central consistiu em avaliar se o intervalo de tempo entre sessões de interações consecutivas poderia ser utilizado como modulador dinâmico da exploração, por meio do algoritmo *Time-Aware LinBoltzmann*.

9.1 Síntese das respostas às perguntas de pesquisa

Os experimentos e análises conduzidos ao longo do estudo permitiram responder de forma articulada às quatro perguntas de pesquisa:

- **PP1:** Constatou-se que protocolos de avaliação *offline* são sistematicamente enviesados contra a exploração, favorecendo políticas gulosas. Esse viés foi confirmado por experimentos em larga escala e pela publicação no RecSys 2025, reforçando a necessidade de metodologias alternativas.
- **PP2:** Para avaliar o tempo de forma realista, foi necessária a adaptação do simulador *KuaiSim*, incorporando um modelo explícito de retorno inter-sessões. Essa modificação possibilitou que tempo se tornasse parte endógena do contexto observado pelas políticas.
- **PP3:** O **Time-Aware LinBoltzmann** não superou os benchmarks em recompensa média, mas alcançou ganhos expressivos em cobertura e diversidade, mostrando que o tempo pode ampliar a variedade e a descoberta de itens sem comprometer de forma relevante a eficácia imediata.

- ❑ **PP4:** A investigação revelou limitações importantes, como escassez de dados temporais adequados, ausência de protocolos padronizados e dependência de simuladores, mas também evidenciou um potencial significativo para desenvolver métodos de recomendação mais personalizados, adaptativos e condizentes com o ritmo de interação dos usuários.

9.2 Contribuições principais

As principais contribuições deste trabalho podem ser resumidas em quatro eixos:

1. Diagnóstico experimental do viés pró-aprofundamento em avaliações *offline* de MAB lineares.
2. Adaptação do simulador *KuaiSim*, tornando-o capaz de suportar experimentos com MAB contextuais e modelagem temporal explícita.
3. Proposição do *Time-Aware LinBoltzmann*, uma política exploratória adaptada ao tempo de retorno dos usuários.
4. Estabelecimento de protocolos de avaliação *online* que permitem investigar de forma justa o papel do tempo na exploração.

9.3 Implicações e perspectivas futuras

Os resultados obtidos sugerem que a incorporação do tempo como fator de exploração não é uma solução imediata para maximizar recompensas, mas constitui um recurso valioso para enriquecer métricas relacionadas à diversidade, cobertura e retenção. Assim, este trabalho reforça a ideia de que avaliações de sistemas de recomendação devem ir além da acurácia ou da recompensa imediata, incorporando também indicadores de longo prazo.

Diversas oportunidades de aprofundamento permanecem em aberto. Uma primeira perspectiva diz respeito ao desenho de novas funções de temperatura. Através do **Time-Aware LinBoltzmann**, foram exploradas apenas variações lineares e normalizadas de δ_t. Funções mais agressivas, não lineares ou combinadas com outros sinais contextuais podem potencializar o impacto do tempo na exploração. Por exemplo, ajustes baseados em funções exponenciais ou transformações mais sofisticadas podem modular de forma mais sensível o grau de exploração diante de padrões temporais distintos.

Outra direção relevante envolve a integração com modelos não lineares. Todos os experimentos desta dissertação foram conduzidos em políticas lineares, mas há espaço considerável para investigar se a consciência temporal pode ser incorporada a variantes mais complexas, como *deep bandits* ou algoritmos baseados em representações latentes

aprendidas por redes neurais. Essa linha de pesquisa pode revelar interações mais ricas entre tempo, contexto e preferências dos usuários.

Do ponto de vista metodológico, há a necessidade de validar os simuladores utilizados contra dados reais de diferentes domínios. O *KuaiSim*, mesmo após as modificações introduzidas, continua sendo uma abstração (ainda que realista) do conjunto de dados *KuaiRand*. Estudos comparativos entre distribuições de tempo geradas pelo simulador e observadas em sistemas reais poderiam aumentar a confiabilidade da avaliação *time-aware*. Além disso, o desenvolvimento de novos simuladores com modelagem temporal mais variada (por exemplo, incluindo sazonalidades e ciclos de uso) constitui uma linha promissora.

As adaptações realizadas no *KuaiSim* também abrem espaço para a exploração de novas formas de simular o tempo. Até o momento, o intervalo de retorno vem sendo gerado em dias a partir da distribuição geométrica proposta pelos autores originais. Contudo, o simulador agora permite incorporar modelos alternativos capazes de interpretar padrões temporais mais granulares, como retornos em horas ou minutos. Essa flexibilidade possibilita, por exemplo, a aplicação de métodos de regressão treinados sobre dados históricos, de modo a prever o tempo de retorno específico de cada usuário. Tal avanço ampliaria a fidelidade da simulação e permitiria investigações mais detalhadas sobre a interação entre recomendações, engajamento e dinâmica temporal.

Por fim, há um espaço amplo para explorar novas métricas orientadas ao usuário. Enquanto esta dissertação se concentrou em recompensa, cobertura, diversidade e sinais básicos de engajamento, trabalhos futuros poderiam incorporar indicadores de longo prazo, como retenção, satisfação percebida, bem-estar digital ou equidade na exposição de itens. Avaliar como o tempo interage com tais dimensões pode enriquecer a compreensão do impacto real das estratégias exploratórias.

Em conjunto, essas perspectivas indicam que a incorporação do tempo como fator de exploração não se limita a um ajuste incremental em algoritmos existentes, mas abre uma linha de pesquisa mais ampla. O caminho iniciado por este trabalho oferece uma base sobre a qual futuros estudos poderão construir soluções mais sofisticadas, realistas e centradas no usuário.

9.3.1 Contribuições científicas e disseminação

Além das contribuições técnicas destacadas anteriormente, este trabalho também gerou avanços no âmbito científico e de disseminação de resultados:

- **Publicação científica**: parte dos resultados foi consolidada no artigo *"Exploitation Over Exploration: Unmasking the Bias in Linear Bandit Recommender Offline Evaluation"* (PIRES et al., 2025), aceito na trilha de Reprodutibilidade do ACM RecSys 2025.

- **Reprodutibilidade e código aberto**: foi disponibilizado publicamente o repositório `exploit-over-explore`[1], que contém a implementação dos experimentos, protocolos de avaliação e scripts associados ao artigo publicado (avaliação *offline*).
- **Ferramentas de simulação**: foi desenvolvido também o repositório `KuaiBandit`[2], que reúne as adaptações realizadas sobre o simulador KuaiSim para dar suporte a bandits contextuais e modelagem temporal. Este repositório encontra-se em fase final de refatoração e documentação, visando futura disponibilização pública. Além disso, está em preparação um artigo dedicado especificamente a essas extensões do simulador, de modo a detalhar sua arquitetura, funcionalidades e potencial de uso em experimentos de avaliação *online*.

[1] <https://github.com/UFSCar-LaSID/exploit-over-explore>
[2] <https://github.com/RecSys-UFSCar/KuaiBandit>

Referências

ADOMAVICIUS, G.; TUZHILIN, A. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. **IEEE Transactions on Knowledge and Data Engineering**, v. 17, n. 6, p. 734–749, 2005.

AGARWAL, A.; MCMAHAN, H. B.; XU, Z. An empirical evaluation of federated contextual bandit algorithms. **arXiv preprint arXiv:2303.10218**, 2023.

AGRAWAL, S.; GOYAL, N. Thompson sampling for contextual bandits with linear payoffs. In: **Proceedings of the 30th International Conference on Machine Learning**. New York, NY, USA: JMLR.org, 2013. (ICML'13), p. 1220–1228.

AKKER, B. van den et al. Practical bandits: An industry perspective. In: **Proceedings of the 17th ACM International Conference on Web Search and Data Mining**. New York, NY, USA: Association for Computing Machinery, 2024. (WSDM '24), p. 1132–1135. ISBN 9798400703713. Disponível em: <https://doi.org/10.1145/3616855.3636449>.

AMEKO, M. K. et al. Offline contextual multi-armed bandits for mobile health interventions: A case study on emotion regulation. In: **Proceedings of the 14th ACM Conference on Recommender Systems**. New York, NY, USA: Association for Computing Machinery, 2020. (RecSys '20), p. 249–258. ISBN 9781450375832. Disponível em: <https://doi.org/10.1145/3383313.3412244>.

AUER, P.; CESA-BIANCHI, N.; FISCHER, P. Finite-time analysis of the multiarmed bandit problem. **Machine Learning**, Springer, v. 47, n. 2, p. 235–256, May 2002. Disponível em: <https://doi.org/10.1023/A:1013689704352>.

BAN, Y.; HE, J.; COOK, C. B. Multi-facet contextual bandits: A neural network perspective. In: **Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining**. New York, NY, USA: Association for Computing Machinery, 2021. (KDD '21), p. 35–45. ISBN 9781450383325. Disponível em: <https://doi.org/10.1145/3447548.3467299>.

BAO, J.; ZHANG, Y. Time-aware recommender system via continuous-time modeling. In: **Proceedings of the 30th ACM International Conference on Information & Knowledge Management**. New York, NY, USA: Association for Computing Machinery, 2021. (CIKM '21), p. 2872–2876. ISBN 9781450384469. Disponível em: <https://doi.org/10.1145/3459637.3482202>.

BASTANI, H.; BAYATI, M.; KHOSRAVI, K. Mostly exploration-free algorithms for contextual bandits. **Management Science**, INFORMS, v. 67, n. 3, p. 1329–1349, 2020. Disponível em: <https://doi.org/10.1287/mnsc.2020.3605>.

BETELLO, F. et al. A reproducible analysis of sequential recommender systems. **IEEE Access**, v. 13, p. 5762–5772, 2025.

BIETTI, A.; AGARWAL, A.; LANGFORD, J. A contextual bandit bake-off. **Journal of Machine Learning Research**, v. 22, n. 133, p. 1–49, 2021. Disponível em: <http://jmlr.org/papers/v22/18-863.html>.

BOBADILLA, J. et al. Recommender systems survey. **Knowledge-Based Systems**, v. 46, p. 109–132, 2013. ISSN 0950-7051. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0950705113001044>.

BOGINA, V. et al. Considering temporal aspects in recommender systems: a survey. **User Modeling and User-Adapted Interaction**, Springer, v. 33, n. 1, p. 81–119, 2023. Disponível em: <https://doi.org/10.1007/s11257-022-09335-w>.

BOUNEFFOUF, D.; RISH, I.; AGGARWAL, C. C. Survey on applications of multi-armed and contextual bandits. In: **Proceedings of the 2020 IEEE Congress on Evolutionary Computation (CEC)**. Piscataway, NJ, USA: IEEE, 2020. p. 1–8.

CAMPOS, P. G.; DÍEZ, F.; CANTADOR, I. Time-aware recommender systems: a comprehensive survey and analysis of existing evaluation protocols. **User Modeling and User-Adapted Interaction**, Springer, v. 24, n. 1, p. 67–119, 2014.

CAÑAMARES, R.; REDONDO, M.; CASTELLS, P. Multi-armed recommender system bandit ensembles. In: **Proceedings of the 13th ACM Conference on Recommender Systems**. New York, NY, USA: Association for Computing Machinery, 2019. (RecSys '19), p. 432–436. ISBN 9781450362436. Disponível em: <https://doi.org/10.1145/3298689.3346984>.

CAO, Z. et al. Privacy matters: Vertical federated linear contextual bandits for privacy protected recommendation. In: **Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining**. New York, NY, USA: Association for Computing Machinery, 2023. (KDD '23), p. 154–166. ISBN 9798400701030. Disponível em: <https://doi.org/10.1145/3580305.3599475>.

CARON, S.; BHAGAT, S. Mixing bandits: a recipe for improved cold-start recommendations in a social network. In: **Proceedings of the 7th Workshop on Social Network Mining and Analysis**. New York, NY, USA: Association for Computing Machinery, 2013. (SNAKDD'13), p. 1–9.

CASTELLS, P.; HURLEY, N.; VARGAS, S. Novelty and diversity in recommender systems. In: RICCI, F.; ROKACH, L.; SHAPIRA, B. (Ed.). **Recommender Systems Handbook**. New York, NY: Springer US, 2022. p. 603–646. ISBN 978-1-0716-2197-4. Disponível em: <https://doi.org/10.1007/978-1-0716-2197-4_16>.

CELIS, L. E. et al. Controlling polarization in personalization: An algorithmic framework. In: **Proceedings of the Conference on Fairness, Accountability, and Transparency**. New York, NY, USA: Association for Computing Machinery, 2019. (FAT'19), p. 160–169.

CESA-BIANCHI, N.; GENTILE, C.; ZAPPELLA, G. A gang of bandits. In: **Proceedings of the 27th International Conference on Neural Information Processing Systems**. Red Hook, NY, USA: Curran Associates Inc., 2013. (NIPS'13), p. 737–745.

CHAPELLE, O.; LI, L. An empirical evaluation of thompson sampling. In: **Proceedings of the 25th International Conference on Neural Information Processing Systems**. Red Hook, NY, USA: Curran Associates Inc., 2011. (NIPS'11), p. 2249–2257.

CHEN, L.; XU, J.; LU, Z. Contextual combinatorial multi-armed bandits with volatile arms and submodular reward. In: **Proceedings of the 32nd Conference on Neural Information Processing Systems**. Red Hook, NY, USA: Curran Associates, Inc., 2018. (NeurIPS'18), p. 3251–3260.

CHEN, S. et al. A unified framework of policy learning for contextual bandit with confounding bias and missing observations. **arXiv preprint arXiv:2303.11187**, 2023.

CHEN, X. et al. Knowledge-guided deep reinforcement learning for interactive recommendation. In: **Proceedings of the 2020 International Joint Conference on Neural Networks**. New York, NY, USA: IEEE, 2020. (IJCNN'20), p. 1–8.

Chen, Yizhe. Contextual bandits to increase user prediction accuracy in movie recommendation system. **ITM Web Conf.**, v. 73, p. 01018, 2025. Disponível em: <https://doi.org/10.1051/itmconf/20257301018>.

DACREMA, M. F.; CREMONESI, P.; JANNACH, D. Are we really making much progress? a worrying analysis of recent neural recommendation approaches. In: **Proceedings of the 13th ACM Conference on Recommender Systems**. New York, NY, USA: Association for Computing Machinery, 2019. (RecSys '19), p. 101–109. ISBN 9781450362436. Disponível em: <https://doi.org/10.1145/3298689.3347058>.

DEREVENTSOV, A.; BIBIN, A. Simulated contextual bandits for personalization tasks from recommendation datasets. In: **Proceedings of the 2022 IEEE International Conference on Data Mining Workshops (ICDMW)**. Piscataway, NJ, USA: IEEE, 2022. p. 1–6. Event: IEEE International Conference on Data Mining Workshops (ICDMW 2022), Orlando, FL, USA, 28 Nov.–1 Dec. 2022.

DU, N. et al. Time-sensitive recommendation from recurrent user activities. In: CORTES, C. et al. (Ed.). **Advances in Neural Information Processing Systems**. Curran Associates, Inc., 2015. v. 28. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2015/file/136f951362dab62e64eb8e841183c2a9-Paper.pdf>.

DUDIK, M.; LANGFORD, J.; LI, L. Doubly robust policy evaluation and learning. In: **Proceedings of the 28th International Conference on International Conference on Machine Learning**. Madison, WI, USA: Omnipress, 2011. (ICML'11), p. 1097–1104.

FAN, Z. et al. Continuous-time sequential recommendation with temporal graph collaborative transformer. In: **Proceedings of the 30th ACM International Conference on Information & Knowledge Management**. New York, NY, USA: Association for Computing Machinery, 2021. (CIKM '21), p. 433–442. ISBN 9781450384469. Disponível em: <https://doi.org/10.1145/3459637.3482242>.

FILIPOVIC, M. et al. Modeling online behavior in recommender systems: The importance of temporal context. In: **Proceedings of the Perspectives on the Evaluation of Recommender Systems Workshop (PERSPECTIVES '21)**. Online: CEUR Workshop Proceedings, 2021. v. 2955. Disponível em: <http://ceur-ws.org/Vol-2955/paper4.pdf>.

FREEMAN, J.; RAWSON, M. Top-k ranking deep contextual bandits for information selection systems. In: **Proceedings of the 2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)**. Piscataway, NJ, USA: IEEE, 2021. p. 2209–2214. Event: IEEE International Conference on Systems, Man, and Cybernetics (SMC 2021), Virtual Conference, 17–20 Oct. 2021.

GAN, M.; KWON, O.-C. A knowledge-enhanced contextual bandit approach for personalized recommendation in dynamic domains. **Knowledge-Based Systems**, v. 251, p. 109158, 2022. ISSN 0950-7051. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0950705122005767>.

GANGAN, E.; KUDUS, M.; ILYUSHIN, E. Survey of multi-armed bandit algorithms applied to recommendation systems. **International Journal of Open Information Technologies**, Laboratory of Open Information Technologies, Lomonosov Moscow State University, v. 9, n. 4, p. 12–27, 2021. ISSN 2307-8162. Disponível em: <https://injoit.org/index.php/j1/article/view/1066>.

GARCELON, E. et al. Improved algorithms for conservative exploration in bandits. In: **Proceedings of the AAAI Conference on Artificial Intelligence**. AAAI Press, 2020. v. 34, n. 4, p. 3962–3969. Disponível em: <https://ojs.aaai.org/index.php/AAAI/article/view/5812>.

GHOORCHIAN, S.; KORTUKOV, E.; MAGHSUDI, S. Non-stationary linear bandits with dimensionality reduction for large-scale recommender systems. **IEEE Open Journal of Signal Processing**, IEEE Computer Society, New York, NY, USA, v. 5, p. 548–558, 2024.

GILOTTE, A. et al. Offline A/B testing for recommender systems. In: **Proceedings of the 11th ACM International Conference on Web Search and Data Mining**. New York, NY, USA: Association for Computing Machinery, 2018. (WSDM'18), p. 198–206.

GOLDBERG, D. et al. Using collaborative filtering to weave an information tapestry. **Commun. ACM**, Association for Computing Machinery, New York, NY, USA, v. 35, n. 12, p. 61–70, dez. 1992. ISSN 0001-0782. Disponível em: <https://doi.org/10.1145/138859.138867>.

GU, H. et al. Robust and efficient algorithms for conversational contextual bandit. **Information Sciences**, v. 657, p. 119993, 2024. ISSN 0020-0255. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0020025523015785>.

GUO, X.; WANG, X.; LIU, X. Adalinucb: Opportunistic learning for contextual bandits. In: **Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI'19)**. Palo Alto, CA, USA: AAAI Press, 2019. p. 2420–2427. ISBN 978-0-9992411-4-1. DOI: 10.24963/ijcai.2019/335. Disponível em: <https://doi.org/10.24963/ijcai.2019/335>.

GUPTA, S. et al. Optimal baseline corrections for off-policy contextual bandits. In: **Proceedings of the 18th ACM Conference on Recommender Systems**. New York, NY, USA: Association for Computing Machinery, 2024. (RecSys'24), p. 722–732.

GUTOWSKI, N. et al. Context enhancement for linear contextual multi-armed bandits. In: **Proceedings of the 2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI)**. Piscataway, NJ, USA: IEEE, 2018. p. 1048–1055. Event: IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI 2018), Volos, Greece, 5–7 Nov. 2018.

HAN, Y. Comparative evaluation, challenges, and diverse applications of multi-armed bandit algorithms. **Highlights in Science, Engineering and Technology**, DR Press, v. 94, p. 206–210, April 2024. Disponível em: <https://drpress.org/ojs/index.php/HSET/article/view/20576>.

HAO, B.; LATTIMORE, T.; SZEPESVÁRI, C. Adaptive exploration in linear contextual bandit. In: CHIAPPA, S.; CALANDRA, R. (Ed.). **Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics (AISTATS 2020)**. Palermo, Italy (Virtual Conference): PMLR, 2020. (Proceedings of Machine Learning Research, v. 108), p. 3536–3545. Disponível em: <https://proceedings.mlr.press/v108/hao20b.html>.

HARIRI, N.; MOBASHER, B.; BURKE, R. Adapting to user preference changes in interactive recommendation. In: **Proceedings of the 24th International Conference on Artificial Intelligence**. Palo Alto, CA, USA: AAAI Press, 2015. (IJCAI'15), p. 4268–4274.

HE, X. et al. Contextual user browsing bandits for large-scale online mobile recommendation. In: **Proceedings of the 14th ACM Conference on Recommender Systems**. New York, NY, USA: Association for Computing Machinery, 2020. (RecSys '20), p. 63–72. ISBN 9781450375832. Disponível em: <https://doi.org/10.1145/3383313.3412234>.

HEJAZINIA, M. et al. Accelerated learning from recommender systems using multi-armed bandit. **arXiv preprint arXiv:1908.06158**, August 2019. Version 1, submitted on 14 Aug 2019. Disponível em: <https://arxiv.org/abs/1908.06158>.

HERLOCKER, J. L. et al. Evaluating collaborative filtering recommender systems. **ACM Trans. Inf. Syst.**, Association for Computing Machinery, New York, NY, USA, v. 22, n. 1, p. 5–53, jan. 2004. ISSN 1046-8188. Disponível em: <https://doi.org/10.1145/963770.963772>.

HU, Y.; KOREN, Y.; VOLINSKY, C. Collaborative filtering for implicit feedback datasets. In: **Proceedings of the 8th IEEE International Conference on Data Mining**. New York, NY, USA: IEEE Computer Society, 2008. (ICDM'08), p. 263–272.

HUANG, J. et al. Keeping dataset biases out of the simulation: A debiased simulator for reinforcement learning based recommender systems. In: **Proceedings of the 14th ACM Conference on Recommender Systems**. New York, NY, USA: Association for Computing Machinery, 2020. (RecSys '20), p. 190–199. ISBN 9781450375832. Disponível em: <https://doi.org/10.1145/3383313.3412252>.

HUANG, L. et al. Position-enhanced and time-aware graph convolutional network for sequential recommendations. **ACM Trans. Inf. Syst.**, Association for Computing Machinery, New York, NY, USA, v. 41, n. 1, jan. 2023. ISSN 1046-8188. Disponível em: <https://doi.org/10.1145/3511700>.

HUANG, W. et al. Fairness-aware bandit-based recommendation. In: **Proceedings of the 2021 IEEE International Conference on Big Data (BigData)**. Piscataway, NJ, USA: IEEE, 2021. p. 1273–1278. Event: IEEE International Conference on Big Data (BigData 2021), Orlando, FL, USA, 15–18 Dec. 2021.

_____. Achieving user-side fairness in contextual bandits. **Human-Centric Intelligent Systems**, Springer, v. 2, n. 1, p. 67–81, 2022. Disponível em: <https://doi.org/10.1007/s44230-022-00008-w>.

HUANG, W.; WU, X. Robustly improving bandit algorithms with confounded and selection biased offline data: a causal approach. In: **Proceedings of the AAAI Conference on Artificial Intelligence**. [s.n.], 2024. Disponível em: <https://ojs.aaai.org/index.php/AAAI/article/view/30027>.

IE, E. et al. Recsim: A configurable simulation platform for recommender systems. **arXiv preprint arXiv:1909.04847**, 2019.

JAGERMAN, R.; MARKOV, I.; RIJKE, M. de. When people change their mind: Off-policy evaluation in non-stationary recommendation environments. In: **Proceedings of the 12th ACM International Conference on Web Search and Data Mining**. New York, NY, USA: Association for Computing Machinery, 2019. (WSDM'19), p. 447–455.

JEUNEN, O.; GOETHALS, B. Top-k contextual bandits with equity of exposure. In: **Proceedings of the 15th ACM Conference on Recommender Systems**. New York, NY, USA: Association for Computing Machinery, 2021. (RecSys '21), p. 310–320. ISBN 9781450384582. Disponível em: <https://doi.org/10.1145/3460231.3474248>.

JIANG, T.; ZENG, J. Time-aware explainable recommendation via updating enabled online prediction. **Entropy**, MDPI, v. 24, n. 11, p. 1639, 2022. Disponível em: <https://doi.org/10.3390/e24111639>.

KADIOĞLU, S.; KLEYNHANS, B. Building higher-order abstractions from the components of recommender systems. In: **Proceedings of the AAAI Conference on Artificial Intelligence**. Washington, DC, USA: AAAI Press, 2024. (AAAI-24), p. 22998–23004.

KHOSRAVI, H. et al. Lnucb-ta: Linear-nonlinear hybrid bandit learning with temporal attention. **arXiv preprint arXiv:2503.00387**, 2025.

KIYOHARA, H.; NOMURA, M.; SAITO, Y. Off-policy evaluation of slate bandit policies via optimizing abstraction. In: **Proceedings of the ACM Web Conference 2024**. New York, NY, USA: Association for Computing Machinery, 2024. (WWW '24), p. 3150–3161. ISBN 9798400701719. Disponível em: <https://doi.org/10.1145/3589334.3645343>.

KRAUTH, K. et al. Do offline metrics predict online performance in recommender systems? **arXiv preprint arXiv:2011.07931**, 2020.

LACERDA, A. Contextual bandits for multi-objective recommender systems. In: **Proceedings of the 2015 Brazilian Conference on Intelligent Systems (BRACIS)**. Piscataway, NJ, USA: IEEE, 2015. p. 68–73. Event: Brazilian Conference on Intelligent Systems (BRACIS 2015), Natal, Brazil, 4–7 Nov. 2015.

_____. Multi-objective ranked bandits for recommender systems. **Neurocomputing**, v. 246, p. 12–24, 2017. ISSN 0925-2312. Brazilian Conference on Intelligent Systems 2015. Disponível em: <https://www.sciencedirect.com/science/article/pii/S092523121730228X>.

LANGFORD, J.; STREHL, A.; WORTMAN, J. Exploration scavenging. In: **Proceedings of the 25th International Conference on Machine Learning (ICML '08)**. New York, NY, USA: Association for Computing Machinery, 2008. (ICML '08), p. 528–535. ISBN 9781605582054. Disponível em: <https://doi.org/10.1145/1390156.1390223>.

LATTIMORE, T.; SZEPESVÁRI, C. **Bandit Algorithms**. Cambridge, United Kingdom: Cambridge University Press, 2020. ISBN 9781108571401. Disponível em: <https://www.cambridge.org/core/books/bandit-algorithms/8E39FD004E6CE036680F90DD0C6F09FC>.

LEE, M.; SIEDAHMED, A.; HEFFERNAN, N. Expert features for a student support recommendation contextual bandit algorithm. In: **Proceedings of the 14th International Conference on Learning Analytics Knowledge (LAK '24)**. New York, NY, USA: Association for Computing Machinery, 2024. (LAK '24), p. 864–870. ISBN 9798400716188. Disponível em: <https://doi.org/10.1145/3636555.3636909>.

LEQI, L. et al. A field test of bandit algorithms for recommendations: Understanding the validity of assumptions on human preferences in multi-armed bandits. In: **Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems**. New York, NY, USA: Association for Computing Machinery, 2023. (CHI '23). ISBN 9781450394215. Disponível em: <https://doi.org/10.1145/3544548.3580670>.

LETARD, A. et al. Bandit algorithms: A comprehensive review and their dynamic selection from a portfolio for multicriteria top-k recommendation. **Expert Systems with Applications**, v. 246, p. 123151, 2024. ISSN 0957-4174. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0957417424000162>.

LI, L. et al. A contextual-bandit approach to personalized news article recommendation. In: **Proceedings of the 19th International Conference on World Wide Web**. New York, NY, USA: Association for Computing Machinery, 2010. (WWW'09), p. 661–670.

_____. An unbiased offline evaluation of contextual bandit algorithms with generalized linear models. In: **Proceedings of the 2011 International Conference on On-line Trading of Exploration and Exploitation 2**. New York, NY, USA: JMLR.org, 2011. (OTEAE'11), p. 19–36.

LI, X. et al. Time–frequency sensitive prompt tuning framework for session-based recommendation. **Expert Systems with Applications**, v. 270, p. 126501, 2025. ISSN 0957-4174. Disponível em: <https://www.sciencedirect.com/science/article/pii/S095741742500123X>.

LIU, B. et al. Transferable contextual bandit for cross-domain recommendation. In: **Proceedings of the 32nd AAAI Conference on Artificial Intelligence**. Palo Alto, CA, USA: AAAI Press, 2018. (AAAI'18), p. 3619–3626.

MA, Y.; WANG, Y.-X.; NARAYANASWAMY, B. Imitation-regularized offline learning. In: CHAUDHURI, K.; SUGIYAMA, M. (Ed.). **Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics**. PMLR, 2019. (Proceedings of Machine Learning Research, v. 89), p. 2956–2965. Disponível em: <https://proceedings.mlr.press/v89/ma19b.html>.

MARY, J.; PREUX, P.; NICOL, O. Improving offline evaluation of contextual bandit algorithms via bootstrapping techniques. In: XING, E. P.; JEBARA, T. (Ed.). **Proceedings of the 31st International Conference on Machine Learning**. Bejing, China: PMLR, 2014. (Proceedings of Machine Learning Research, 2), p. 172–180. Disponível em: <https://proceedings.mlr.press/v32/mary14.html>.

MCINERNEY, J. et al. Explore, exploit, and explain: personalizing explainable recommendations with bandits. In: **Proceedings of the 12th ACM Conference on Recommender Systems**. New York, NY, USA: Association for Computing Machinery, 2018. (RecSys '18), p. 31–39. ISBN 9781450359016. Disponível em: <https://doi.org/10.1145/3240323.3240354>.

MCNEE, S. M.; RIEDL, J.; KONSTAN, J. A. Being accurate is not enough: how accuracy metrics have hurt recommender systems. In: **CHI '06 Extended Abstracts on Human Factors in Computing Systems**. New York, NY, USA: Association for Computing Machinery, 2006. (CHI EA '06), p. 1097–1101. ISBN 1595932984. Disponível em: <https://doi.org/10.1145/1125451.1125659>.

MEHROTRA, R.; XUE, N.; LALMAS, M. Bandit based optimization of multiple objectives on a music streaming platform. In: **Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining**. New York, NY, USA: Association for Computing Machinery, 2020. (KDD '20), p. 3224–3233. ISBN 9781450379984. Disponível em: <https://doi.org/10.1145/3394486.3403374>.

MEZNI, H.; FAYALA, M. Time-aware service recommendation: Taxonomy, review, and challenges. **Software: Practice and Experience**, v. 48, n. 11, p. 2080–2108, 2018. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/spe.2605>.

MLADENOV, M. et al. Recsim ng: toward principled uncertainty modeling for recommender ecosystems.(2021). **arXiv preprint arXiv:2103.08057**, 2021.

MU, T. et al. Factored dro: Factored distributionally robust policies for contextual bandits. In: **Advances in Neural Information Processing Systems (NeurIPS 2022)**. New Orleans, LA, USA: Curran Associates, Inc., 2022.

NGUYEN, D. et al. Adaptive sequence learning: Contextual multi-armed bandit approach. In: **2023 IEEE International Conference on Dependable, Autonomic and Secure Computing; International Conference on Pervasive Intelligence and Computing; International Conference on Cloud and Big Data Computing; International Conference on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)**. Exeter, United Kingdom: IEEE, 2023.

NGUYEN, T. T.; LAUW, H. W. Dynamic clustering of contextual multi-armed bandits. In: **Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management**. New York, NY, USA: Association for Computing Machinery, 2014. (CIKM '14), p. 1959–1962. ISBN 9781450325981. Disponível em: <https://doi.org/10.1145/2661829.2662063>.

PARK, H.; FARADONBEH, M. K. S. Analysis of thompson sampling for partially observable contextual multi-armed bandits. **IEEE Control Systems Letters**, v. 6, p. 2150–2155, 2022.

PENG, T. et al. Tagrec: Temporal-aware graph contrastive learning with theoretical augmentation for sequential recommendation. **IEEE Transactions on Knowledge and Data Engineering**, v. 37, n. 5, p. 3015–3029, 2025.

PIRES, P. R. et al. Exploitation over exploration: Unmasking the bias in linear bandit recommender offline evaluation. In: **Proceedings of the Nineteenth ACM Conference on Recommender Systems**. New York, NY, USA: Association for Computing Machinery, 2025. (RecSys '25), p. 736–745. ISBN 9798400713644. Disponível em: <https://doi.org/10.1145/3705328.3748166>.

QASSIMI, S.; RAKRAK, S. Multi-objective contextual bandits in recommendation systems for smart tourism. **Scientific Reports**, v. 15, n. 1, p. 13669, 2025. Disponível em: <https://doi.org/10.1038/s41598-025-89920-2>.

RABIU, I. et al. Recommender system based on temporal models: A systematic review. **Applied Sciences**, MDPI, v. 10, n. 7, p. 2204, 2020. Disponível em: <https://www.mdpi.com/2076-3417/10/7/2204>.

RAO, D. Contextual bandits for adapting to changing user preferences over time. **arXiv preprint arXiv:2009.10073**, p. 1–11, 2020.

RESNICK, P.; VARIAN, H. R. Recommender systems. **Commun. ACM**, Association for Computing Machinery, New York, NY, USA, v. 40, n. 3, p. 56–58, mar. 1997. ISSN 0001-0782. Disponível em: <https://doi.org/10.1145/245108.245121>.

RICCI, F.; ROKACH, L.; SHAPIRA, B. **Recommender Systems Handbook**. 2nd. ed. Boston, MA: Springer, 2015.

ROHDE, D. et al. Recogym: A reinforcement learning environment for the problem of product recommendation in online advertising. **arXiv preprint arXiv:1808.00720**, 2018.

SACHDEVA, N.; SU, Y.; JOACHIMS, T. Off-policy bandits with deficient support. In: **Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '20)**. Virtual Event, USA: Association for Computing Machinery (ACM), 2020. p. 965–975.

SAITO, Y. et al. Open Bandit Dataset and Pipeline: Towards realistic and reproducible off-policy evaluation. In: **Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks**. Red Hook, NY, USA: Curran Associates Inc., 2021. (NeurIPS'21), p. 1–14.

SAITO, Y.; JOACHIMS, T. Off-policy evaluation for large action spaces via embeddings. In: **Proceedings of the 39th International Conference on Machine Learning (ICML 2022)**. Baltimore, MD, USA: PMLR, 2022. (Proceedings of Machine Learning Research, v. 162), p. 19089–19122. Disponível em: <https://proceedings.mlr.press/v162/saito22a.html>.

SAITO, Y.; REN, Q.; JOACHIMS, T. Off-policy evaluation for large action spaces via conjunct effect modeling. In: **Proceedings of the 40th International Conference on Machine Learning (ICML 2023)**. Honolulu, HI, USA: PMLR, 2023. (Proceedings of Machine Learning Research, v. 202), p. 29683–29701. Disponível em: <https://proceedings.mlr.press/v202/saito23a.html>.

SARWAR, B. et al. Item-based collaborative filtering recommendation algorithms. In: **Proceedings of the 10th International Conference on World Wide Web (WWW '01)**. Hong Kong, China: Association for Computing Machinery (ACM), 2001. p. 285–295.

SCHEIDT, T.; BEEL, J. Time-dependent evaluation of recommender systems. In: **Proceedings of the Perspectives on the Evaluation of Recommender Systems Workshop (PERSPECTIVES'21), co-located with the 15th ACM Conference on Recommender Systems (RecSys 2021)**. Amsterdam, The Netherlands: CEUR-WS.org, 2021. (CEUR Workshop Proceedings, v. 2955), p. 1–7. Disponível em: <https://ceur-ws.org/Vol-2955/paper10.pdf>.

SHANI, G.; GUNAWARDANA, A. Evaluating recommendation systems. In: RICCI, F. et al. (Ed.). **Recommender Systems Handbook**. New York, NY, USA: Springer US, 2011. cap. 8, p. 257–259. ISBN 978-0-387-85819-7.

SHEN, C. et al. Hyperbandit: Contextual bandit with hypernetwork for time-varying user preferences in streaming recommendation. In: **Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23)**. Birmingham, United Kingdom: Association for Computing Machinery (ACM), 2023. p. 2274–2283.

SHERMAN, J.; MORRISON, W. J. Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. **The Annals of Mathematical Statistics**, Institute of Mathematical Statistics, v. 21, n. 1, p. 124–127, 1950.

SHI, J.-C. et al. Virtual-taobao: Virtualizing real-world online retail environment for reinforcement learning. In: **Proceedings of the AAAI Conference on Artificial Intelligence**. Honolulu, HI, USA: AAAI Press, 2019. v. 33, n. 1, p. 4902–4909.

SHI, Q. et al. Deep neural network with linucb: A contextual bandit approach for personalized recommendation. In: **Companion Proceedings of the ACM Web Conference 2023 (WWW '23 Companion)**. Austin, TX, USA: Association for Computing Machinery (ACM), 2023. p. 129–137.

SHIMIZU, T. et al. Effective off-policy evaluation and learning in contextual combinatorial bandits. In: **Proceedings of the 18th ACM Conference on Recommender Systems (RecSys '24)**. Bari, Italy: Association for Computing Machinery (ACM), 2024. p. 733–741.

SILVA, N. et al. Multi-armed bandits in recommendation systems: A survey of the state-of-the-art and future directions. **Expert Systems with Applications**, Elsevier Science Publishers B. V., Amsterdam, Netherlands, v. 197, n. 1, p. 1–17, 2022.

SONG, L.; FRAGOULI, C.; SHAH, D. Interactions between learning and broadcasting in wireless recommendation systems. In: **Proceedings of the 2019 IEEE International Symposium on Information Theory**. New York, NY, USA: IEEE, 2019. (ISIT'19), p. 2549–2553.

STAVINOVA, E. et al. Synthetic data-based simulators for recommender systems: A survey. **arXiv preprint arXiv:2206.11338**, 2022.

SUTTON, R. S.; BARTO, A. G. **Reinforcement Learning: An Introduction**. 1st. ed. Cambridge, MA, USA: MIT Press, 1998. ISBN 9780262039246.

SWAMINATHAN, A.; JOACHIMS, T. Counterfactual risk minimization: Learning from logged bandit feedback. In: **Proceedings of the 32nd International Conference on Machine Learning (ICML 2015)**. Lille, France: PMLR, 2015. p. 814–823. Disponível em: <https://proceedings.mlr.press/v37/swaminathan15.html>.

SáNCHEZ, P.; BELLOGíN, A. Time and sequence awareness in similarity metrics for recommendation. **Information Processing Management**, v. 57, n. 3, p. 102228, 2020. ISSN 0306-4573. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0306457319307678>.

TAKEMORI, S. et al. Submodular bandit problem under multiple constraints. In: **Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence**. New York, NY, USA: JMLR.org, 2020. (UAI'20), p. 191–200.

TEKIN, C.; TURğAY, E. Multi-objective contextual multi-armed bandit with a dominant objective. **IEEE Transactions on Signal Processing**, v. 66, n. 14, p. 3799–3813, 2018.

TRACà, S.; RUDIN, C.; YAN, W. Reducing exploration of dying arms in mortal bandits. In: **Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence**. New York, NY, USA: JMLR.org, 2019. (UAI'19), p. 156–163.

TRAN-THE, H. et al. Combining online learning and offline learning for contextual bandits with deficient support. **arXiv preprint arXiv:2107.11533**, 2021.

TUCKER, A.; JOACHIMS, T. Variance-minimizing augmentation logging for counterfactual evaluation in contextual bandits. In: **Proceedings of the 16th ACM International Conference on Web Search and Data Mining (WSDM '23)**. Singapore: Association for Computing Machinery (ACM), 2023. p. 1111–1119.

VARATHARAJAH, Y.; BERRY, B. A contextual-bandit-based approach for informed decision-making in clinical trials. **Life**, MDPI, v. 12, n. 8, p. 1277, 2022. Disponível em: <https://www.mdpi.com/2075-1729/12/8/1277>.

WANG, C.; SHI, L.; LUO, J. Adaptive noise exploration for neural contextual multi-armed bandits. **Algorithms**, MDPI, v. 18, n. 2, p. 56, 2025. Disponível em: <https://www.mdpi.com/1999-4893/18/2/56>.

WANG, C.; WANG, K.; HE, X. Biucb: A contextual bandit algorithm for cold-start and diversified recommendation. In: **2017 IEEE International Conference on Big Knowledge (ICBK)**. Hefei, China: IEEE, 2017. p. 115–122.

WANG, H.; WU, Q.; WANG, H. Learning hidden features for contextual bandits. In: **Proceedings of the 25th ACM International on Conference on Information and Knowledge Management**. New York, NY, USA: Association for Computing Machinery, 2016. (CIKM'16), p. 1633–1642.

_____. Factorization bandits for interactive recommendation. In: **Proceedings of the 31st AAAI Conference on Artificial Intelligence**. Palo Alto, CA, USA: AAAI Press, 2017. (AAAI'17), p. 2695–2702.

WANG, H.; ZARIPHOPOULOU, T.; ZHOU, X. Y. Reinforcement learning in continuous time and space: A stochastic control approach. **Journal of Machine Learning Research**, v. 21, n. 198, p. 1–34, 2020.

WANG, K. et al. Rl4rs: A real-world dataset for reinforcement learning based recommender system. In: **Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23)**. Taipei, Taiwan: Association for Computing Machinery (ACM), 2023. p. 2935–2944.

WANG, Y.-X.; AGARWAL, A.; DUDÍK, M. Optimal and adaptive off-policy evaluation in contextual bandits. In: PRECUP, D.; TEH, Y. W. (Ed.). **Proceedings of the 34th International Conference on Machine Learning**. PMLR, 2017. (Proceedings of Machine Learning Research, v. 70), p. 3589–3597. Disponível em: <https://proceedings.mlr.press/v70/wang17a.html>.

WU, Q.; IYER, N.; WANG, H. Learning contextual bandits in a non-stationary environment. In: **The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval**. New York, NY, USA: Association for Computing Machinery, 2018. (SIGIR '18), p. 495–504. ISBN 9781450356572. Disponível em: <https://doi.org/10.1145/3209978.3210051>.

WU, Q.; LI, Y.; WANG, H. Dynamic ensemble of contextual bandits to satisfy users' changing interests. In: **Proceedings of the World Wide Web Conference (WWW '19)**. San Francisco, CA, USA: Association for Computing Machinery (ACM), 2019. p. 2080–2091.

WU, Q. et al. Returning is believing: Optimizing long-term user engagement in recommender systems. In: **Proceedings of the 2017 ACM on Conference on Information and Knowledge Management**. New York, NY, USA: Association for Computing Machinery, 2017. (CIKM'17), p. 1927–1936.

XIE, H.; TANG, Q.; ZHU, Q. A multiplier bootstrap approach to designing robust algorithms for contextual bandits. **IEEE Transactions on Neural Networks and Learning Systems**, v. 34, n. 12, p. 9887–9899, 2023.

XU, P. et al. Neural contextual bandits with deep representation and shallow exploration. In: **Proceedings of the 10th International Conference on Learning Representations (ICLR 2022)**. Virtual Conference: OpenReview.net, 2022. Disponível em: <https://openreview.net/forum?id=xnYACQquaGV>.

XU, X. et al. Contextual-bandit based personalized recommendation with time-varying user interests. In: **Proceedings of the 34th AAAI Conference on Artificial Intelligence**. Palo Alto, CA, USA: AAAI Press, 2020. (AAAI'20), p. 6518–6525.

YAN, C. et al. Dynamic clustering based contextual combinatorial multi-armed bandit for online recommendation. **Knowledge-Based Systems**, v. 257, p. 109927, 2022. ISSN 0950-7051. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0950705122010206>.

_____. Thompson sampling with time-varying reward for contextual bandits. In: **Database Systems for Advanced Applications (DASFAA 2023)**. Zhangjiajie, China: Springer, 2023. (Lecture Notes in Computer Science, v. 13986), p. 54–63.

YANG, H.; YANG, C. Tignn-rl: Enabling time-sensitive and context-aware intelligent decision-making with dynamic graphs in recommender systems and biomechanics knowledge. **Molecular & Cellular Biomechanics**, Society for Integrative Oncology and Nanomedicine, v. 22, n. 3, p. 1339, February 2025. Disponível em: <https://sin-chn.com/index.php/mcb/article/view/1339>.

YE, W. et al. Time matters: Sequential recommendation with complex temporal information. In: **Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20)**. Virtual Event, China: Association for Computing Machinery (ACM), 2020. p. 1459–1468.

YOU, J. et al. Hierarchical temporal convolutional networks for dynamic recommender systems. In: **The World Wide Web Conference**. New York, NY, USA: Association for Computing Machinery, 2019. (WWW '19), p. 2236–2246. ISBN 9781450366748. Disponível em: <https://doi.org/10.1145/3308558.3313747>.

YU, S. et al. Careforme: Contextual multi-armed bandit recommendation framework for mental health. In: **Proceedings of the 2024 IEEE/ACM 11th International Conference on Mobile Software Engineering and Systems (MOBILESoft '24)**. Lisbon, Portugal: IEEE, 2024. p. 144–154.

ZENG, C. et al. Online context-aware recommendation with time-varying multi-armed bandit. In: **Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)**. San Francisco, CA, USA: Association for Computing Machinery (ACM), 2016. p. 2025–2034.

ZENG, S. et al. Partially observable contextual bandits with linear payoffs. In: **Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2025)**. Hyderabad, India: IEEE, 2025. p. 1–5.

ZHANG, Q. et al. Neural time-aware sequential recommendation by jointly modeling preference dynamics and explicit feature couplings. **IEEE Transactions on Neural Networks and Learning Systems**, v. 33, n. 10, p. 5125–5137, 2022.

ZHANG, S. et al. Deep learning based recommender system: A survey and new perspectives. **ACM Comput. Surv.**, Association for Computing Machinery, New York, NY, USA, v. 52, n. 1, fev. 2019. ISSN 0360-0300. Disponível em: <https://doi.org/10.1145/3285029>.

ZHANG, X. et al. Conversational contextual bandit: Algorithm and application. In: **Proceedings of The Web Conference 2020**. New York, NY, USA: Association for Computing Machinery, 2020. (WWW '20), p. 662–672. ISBN 9781450370233. Disponível em: <https://doi.org/10.1145/3366423.3380148>.

ZHANG, Y. et al. A time-aware self-attention based neural network model for sequential recommendation. **Applied Soft Computing**, v. 133, p. 109894, 2023. ISSN 1568-4946. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1568494622009437>.

ZHAO, K. et al. KuaiSim: a comprehensive simulator for recommender systems. In: **Proceedings of the 37th International Conference on Neural Information Processing Systems**. Red Hook, NY, USA: Curran Associates Inc., 2023. (NIPS'23), p. 44880–44897.

ZHENG, J. et al. Neural contextual combinatorial bandit under non-stationary environment. In: **Proceedings of the 2023 IEEE International Conference on Data Mining (ICDM)**. Shanghai, China: IEEE, 2023. p. 854–863.

ZHOU, C. et al. Conversational music recommendation based on bandits. In: **Proceedings of the 2020 IEEE International Conference on Knowledge Graph**. New York, NY, USA: IEEE, 2020. (ICKG'20), p. 41–48.

ZHOU, S. et al. Interactive recommender system via knowledge graph-enhanced reinforcement learning. In: **Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval**. New York, NY, USA: Association for Computing Machinery, 2020. (SIGIR'20), p. 179–188.

ZHOU, Z.; XU, R.; BLANCHET, J. Learning in generalized linear contextual bandits with stochastic delays. In: WALLACH, H. et al. (Ed.). **Advances in Neural Information Processing Systems**. Curran Associates, Inc., 2019. v. 32. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2019/file/56cb94cb34617aeadff1e79b53f38354-Paper.pdf>.

ZHU, Z. et al. Non-stationary contextual bandit learning via neural predictive ensemble sampling. **arXiv preprint arXiv:2310.07786**, 2023. Disponível em: <https://arxiv.org/abs/2310.07786>.