A Hybrid Meta-Learning and MAB Approach for Context-Specific Multi-Objective Recommendation Optimization

Tiago Cunha[1,*,†], Andrea Marchini[2,†]

[1]Expedia Group, Portugal
[2]Expedia Group, United Kingdom

Abstract

Recommender systems in online marketplaces face the challenge of balancing multiple objectives to satisfy various stakeholders, including customers, providers, and the platform itself. This paper introduces Juggler-MAB, a hybrid approach that combines meta-learning with Multi-Armed Bandits (MAB) to address the limitations of existing multi-stakeholder recommendation systems. Our method extends the Juggler framework, which uses meta-learning to predict optimal weights for utility and compensation adjustments, by incorporating a MAB component for real-time, context-specific refinements. We present a two-stage approach where Juggler provides initial weight predictions, followed by MAB-based adjustments that adapt to rapid changes in user behavior and market conditions. Our system leverages contextual features such as device type and brand to make fine-grained weight adjustments based on specific segments. To evaluate our approach, we developed a simulation framework using a dataset of 0.6 million searches from Expedia's lodging booking platform. Results show that Juggler-MAB outperforms the original Juggler model across all metrics, with NDCG improvements of 2.9%, a 13.7% reduction in regret, and a 9.8% improvement in best arm selection rate.

Keywords

Multi-Stakeholder, Multi-Armed bandits, Meta-Learning

1. Introduction

Recommender systems often focus solely on user satisfaction. However, in many real-world applications, particularly in online marketplaces, multiple stakeholders' interests need to be considered [1, 2]. These stakeholders typically include users (customers), item providers (e.g., hotel owners), and the platform. Multi-stakeholder recommenders aim to balance these diverse and often conflicting objectives [3, 4].

The Juggler framework [5] was introduced to address this multi-stakeholder recommendation problem by using meta-learning [6, 7] to predict optimal weights for utility and compensation adjustments in real-time scoring. Deployed in production, Juggler has been an integral part of the Lodging Ranking stack at Expedia. However, Juggler's reliance on a pre-configured set of five options for relevance and compensation limits its ability to fine-tune recommendations for specific contexts. Additionally, its infrequent training cycles make it less responsive to rapid changes in traffic patterns across segments.

To address these limitations, we propose a two-step approach that combines meta-learning (Juggler) with Multi-Armed Bandits for multi-stakeholder recommendations for real-time weight adjustments. This approach aims to: 1) provide more granular weight adjustments based on specific segments (e.g., device type, brand) and 2) adapt quickly to changes in traffic patterns without requiring frequent retraining of the main Juggler model. Our research questions are:

- Can the integration of MAB with Juggler improve the performance and adaptability of multi-stakeholder recommendations in online marketplaces?
- Are contextual features useful to improve the MAB's effectiveness at making the right decisions?

SURE workshop held in conjunction with the 18th ACM Conference on Recommender Systems (RecSys), 2024, in Bari, Italy.
*Corresponding author.
†These authors contributed equally.
tsacunha@expediagroup.com (T. Cunha); amarchini@expediagroup.com (A. Marchini)

The rest of this paper is organized as follows: Section 2 presents the related work, while Section 3 introduces the proposed hybrid solution. Section 4 covers the experimental setup used to validate the proposal and while Section 5 reports the results to the research questions. Lastly, Section 6 highlights the main conclusions and avenues for future work.

2. Background and Related Work

The Juggler framework [5] was introduced as a meta-learning approach to address the multi-stakeholder recommendation problem. It dynamically predicts the ideal weights for utility (user relevance) and compensation (platform revenue) for each search query. The meta-model leverages a collection of historical search queries and learns the mapping between the search context and the ideal utility and compensation weights, learned via offline simulations. Juggler selects from five pre-configured options, each representing a different balance between relevance and compensation: 1) Lower relevance, lower compensation, 2) Lower relevance, higher compensation, 3) Neutral relevance, neutral compensation, 4) Higher relevance, lower compensation and 5) Higher relevance, higher compensation. The pre-configured options refer to sections of the search space which are explored to identify different directions of improvement, while reducing the number of options to ultimately choose from. It is noteworthy that although the pre-configured options are fixed, the actual instantiation of weights for each option depends on the ranking problem characteristics and Juggler framework hyper-parameters. While Juggler has shown success in production, its reliance on these fixed options and infrequent training cycles limits its adaptability to rapid changes in user behavior and market conditions.

Multi-Armed Bandits (MAB) are a class of reinforcement learning algorithms that balance exploration and exploitation in decision-making processes [8]. In the context of recommenders, MABs have been used to address the exploration-exploitation dilemma and to adapt to changing user preferences [9, 10].

The integration of meta-learning and bandit algorithms has been explored in other domains, such as algorithm selection [11] and hyperparameter optimization [12]. Our work extends these ideas to the realm of multi-stakeholder recommendations, addressing the unique challenges of online marketplaces.

Several studies have addressed the challenge of balancing multiple objectives in recommender systems. Rodriguez et al. [13] proposed a multi-objective optimization approach for job recommendations. Nguyen et al. [14] introduced a multi-objective learning to re-rank approach to optimize online marketplaces for multiple stakeholders. Sürer et al. [15] explored multi-stakeholder recommendation with provider constraints. These approaches provide valuable insights into balancing multiple objectives, but our proposed method aims to extend their capabilities by combining meta-learning with multi-armed bandits for enhanced adaptability in dynamic online marketplaces.

Recent developments in industry have led to the creation of self-service platforms for deploying contextual bandits, such as AdaptEx [16]. These platforms provide powerful tools for optimizing user experiences at scale, which we leverage in our hybrid approach to combine the strengths of meta-learning and MAB algorithms. To evaluate our approach, we utilized a custom simulation framework based on real-world data from an online travel marketplace. This allowed us to assess the performance of our system in a controlled yet realistic setting, similar to other sophisticated simulation environments [17].

3. Juggler with MAB

We present a hybrid approach that combines the Juggler framework's meta-learning capabilities [5] with a MAB system powered by the AdaptEx SDK [16]. This approach, which we call "Juggler-MAB" aims to address the limitations of the original Juggler system while leveraging the adaptive capabilities of contextual bandits. The Juggler-MAB system operates in two stages:

1. **Juggler Stage**: The meta-learning model predicts initial utility and compensation weights based on search context.
2. **MAB Stage**: A contextual MAB refines these weights in real-time based on user interactions and search features.

The Juggler framework selects from five pre-configured options for utility and compensation weights, providing a coarse adjustment of the recommendation strategy based on the search context. These options range from lower relevance and compensation to higher relevance and compensation, as described in [5] and aim to tackle the main issues in multi-objective optimization.

The MAB component introduces fine-grained adjustments to the Juggler-predicted weights. Each arm of the bandit represents a small corrective measure to be applied to the utility and compensation weights to improve relevance.

The key features of our MAB implementation include:

1. **Contextual arms**: The contextual bandits consider contextual features (e.g., device type, brand) when selecting arms.
2. **Reward function**: We use Normalized Discounted Cumulative Gain (NDCG) as a proxy for Conversion Rate, allowing for offline simulation and evaluation.
3. **Exploration strategy**: We employ epsilon-greedy and Thompson Sampling for its ability to balance exploration and exploitation effectively [8].

The integration of Juggler and MAB is achieved through an additive approach in the scoring function:

$$\begin{aligned} sortScore =& (w_{utility}^{Juggler} + w_{utility}^{MAB}) \cdot utilityScore \\ &+ (w_{comp}^{Juggler} + w_{comp}^{MAB}) \cdot compensationScore \end{aligned} \tag{1}$$

where $w_{utility}^{Juggler}$ and $w_{comp}^{Juggler}$ are the weights predicted by Juggler, and $w_{utility}^{MAB}$ and w_{comp}^{MAB} are the corrective weights determined by the MAB.

We formulate our contextual MAB problem as follows: let $\mathcal{A}$ be the set of arms, where each arm $a \in \mathcal{A}$ represents a pair of corrective weights $(w_{utility}^{MAB}, w_{compensation}^{MAB})$. The context $x_t \in \mathcal{X}$ at time t includes features such as device or brand. The reward r_t is defined as the NDCG of the resulting ranking. The goal is to find a policy $\pi : \mathcal{X} \to \mathcal{A}$ that maximizes the expected cumulative reward:

$$\max_{\pi} \mathbb{E}\left[\sum_{t=1}^{T} r_t(x_t, \pi(x_t))\right] \tag{2}$$

where T is the time horizon.

We explored various methods to combine Juggler's predictions with MAB corrections, ultimately settling on the additive approach described above. We carefully selected contextual variables that would help identify under-performing segments in the Juggler model, such as device type and brand. Balancing multiple objectives in a single reward function required careful consideration. We chose NDCG as an initial approach due to its widely accepted usage, with plans to explore more complex multi-objective reward functions in future work.

To evaluate our hybrid approach, we developed a custom simulator that allows us to test various configurations offline using historical data. The simulator, built on Expedia data, enables to:

1. Replay historical searches and user interactions. Data is loaded on a daily basis, consisting of data for each property in each search and the respective user clicks and bookings.
2. Apply the Juggler-MAB model to generate new rankings. The MAB is sampled (ppotentially using contextual data) and the retrieved arm is included in the ranking formula, yielding the simulated score and the final ranking.
3. Evaluate the performance using both immediate (e.g., clicks) and delayed (e.g., bookings) feedback. The reward function evaluates the simulated rankings and information about the arm sampled, reward and contextual information (if any) is provided to the MAB, to update its internal state.

The simulation framework provides a safe environment to test and refine our approach before considering online deployment.

4. Experimental Setup

We used a dataset of 0.6 million searches from Expedia's lodging booking platform, covering a period of 31 consecutive days. The data has over 600000 distinct properties across approximately 41000 distinct destinations, with feedback sparsity over 96%. The dataset includes features such as device type, brand, destination, and historical user interactions.

We compared several variants of the proposed Juggler-MAB hybrid approach against the original Juggler model [5]. We tested several MAB algorithms, ranging from classical (i.e. no contextual features) to contextual bandits:

- Gaussian Thompson (GT): a classical bandit using Thompson Sampling assuming a Gaussian Distribution of reward value.
- ϵ-greedy: a classical bandit using a vanilla implementation of the canonical algorithm. We have used $\epsilon = 0.1$ and $\epsilon = 0.3$.
- Recursive Least Squares with Thompson Sampling (RLS): a contextual bandit using a linear model with a vector of means and a matrix of variances-covariances.

The experiments use the actual production Juggler model predictions for each search. This improves the reliability of Juggler's predictions, which in turn leads to more robust estimates of the MAB's effect. We then implemented the MAB component using the AdaptEx SDK [16], with the following configuration:

- Arm space: we explore 3 different values for each arm, respectively $w_{utility}^{MAB} \in \{-0.3, 0.0, 0.3\}$ and $w_{comp}^{MAB} \in \{-0.2, 0.0, 0.2\}$. The selected weights are determined via domain knowledge, also ensuring non-zero weights.
- Contextual features: several low cardinality categorical search features were tested, with 3 being identified as the most important: brand, user device and geographical categorization of the search destination, e.g. neighborhood vs city.
- Exploration strategy: Thompson Sampling and ϵ-greedy
- Reward: Normalized Discounted Cumulative Gain (NDCG), to determine how well can MAB algorithms correct towards relevance and expected conversion rate improvement.

5. Results and Discussion

Table 1 summarizes the main results. For each bandit, we report the average reward, regret and the percentage of best arm selections across all searches. The best results per metric are highlighted in bold. Notice that regret is best when lowest, the remaining metrics are better when maximized.

Bandit	avg(reward)	avg(regret)	best arm %
Juggler	0.1776	0.0373	0.7515
GT	0.1791	0.0358	0.7866
ϵ-greedy (0.3)	0.1811	0.0339	0.8095
ϵ-greedy (0.1)	0.1824	0.0325	0.8218
RLS_{brand}	**0.1827**	**0.0322**	**0.8252**
RLS_{device}	0.1822	0.0327	0.8200
RLS_{geo}	0.1825	0.0325	0.8228
$RLS_{geo,\ brand}$	**0.1827**	0.0323	0.8246
$RLS_{device,\ brand}$	**0.1827**	**0.0322**	0.8228
$RLS_{geo,\ device}$	**0.1827**	**0.0322**	0.8247
$RLS_{geo,\ device,\ brand}$	0.1826	0.0323	0.8246

Table 1
Aggregated reward, regret and the percentage of best arm selections results for all bandits and baselines.

Our Juggler-MAB hybrid approach outperformed the Juggler baseline across all metrics for all bandits proposed. The NDCG improvements range from +0.8% for GT bandit, all the way to +2.9% in several RLS bandits. In terms of regret, we achieve a reduction of 13.7% and an improvement in best arm selection rate of 9.8%.

The ϵ-greedy algorithms provide very strong baselines, especially when $\epsilon = 0.1$. GT bandit is clearly the worst bandit, but yet useful since it outperforms the baseline. Among the contextual bandits, the best one across all metrics is the RLS_{brand}. Interestingly, when using more contextual features, we did not achieve better performance. Further investigations are required to identify what matters to define the context.

We performed Wilcoxon signed-rank tests and observed no statistical difference between all RLS bandits. The Critical Difference [18] diagram for the remaining bandits is shown in Figure 1. The results show no statistical significant difference between RLS and $\epsilon = 0.1$, hinting that contextual features are not meaningful. However, all RLS bandits are better than the baselines. To note as well how all bandits are better than the Juggler baseline - this is a testament to the value of hybrid approach proposed.

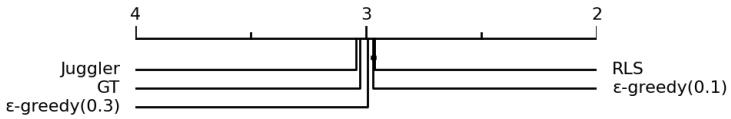

Figure 1: Critical Difference diagram shows superiority of bandits over multiple baselines, including simpler bandits.

Figure 2 shows the learning dynamics for all bandits across all days in the data sample. To improve interpretation, we include only the best contextual bandit. The Juggler-MAB demonstrated fast adaptation to changing conditions. We observed that the MAB component was able to make fine-grained adjustments to the Juggler predictions, resulting in improved performance.

We inspect now Juggler-MAB's effect on lodging ranking top-10 average statistics in Table 2. The results are reported as differences to the Juggler baseline, as we cannot expose the sensitive raw data.

Metric	ϵ-greedy (0.3)	ϵ-greedy (0.1)	RLS
daily price	-0.7278	-0.8324	-0.8595
guest rating	0.0416	0.0572	0.0604
star rating	0.0499	0.0747	0.0796
margin %	-0.0034	-0.0045	-0.0048
margin $	-0.6285	-0.8222	-0.8633

Table 2
Differences in several metrics in top-10 positions.

The results show a clear pattern for all bandits: average daily price decreases and guest and star ratings increase as NDCG improves. On the contrary, margin % and margin $ decreases, which could pose problems to the marketplace objectives and long term health. The expectations, to be validated via AB test, is that the increase in relevance will lead to an improvement in conversion rate which can offset the impact in profit per transaction.

Diving now deeper into the arms selection per bandit, we present Figure 3. The results show a clear and expected preference towards arms lower compensation weights, as they are not aligned with the NDCG reward. However, it is interesting to observe that the best bandit has learned that not only is it ideal to decrease compensation, but also to increase or decrease relevance depending on the context.

Despite the overall positive results, we identified two limitations. First, the reward function considers only a single dimension of the problem (i.e. relevance), which explains the impact to the compensation component. Future work will address this limitation by using multi-objective optimization techniques [13]. Second, our current simulations use historical interactions with a deterministic logging policy, introducing bias. To address it, we will implement off-policy evaluation techniques [19, 20]

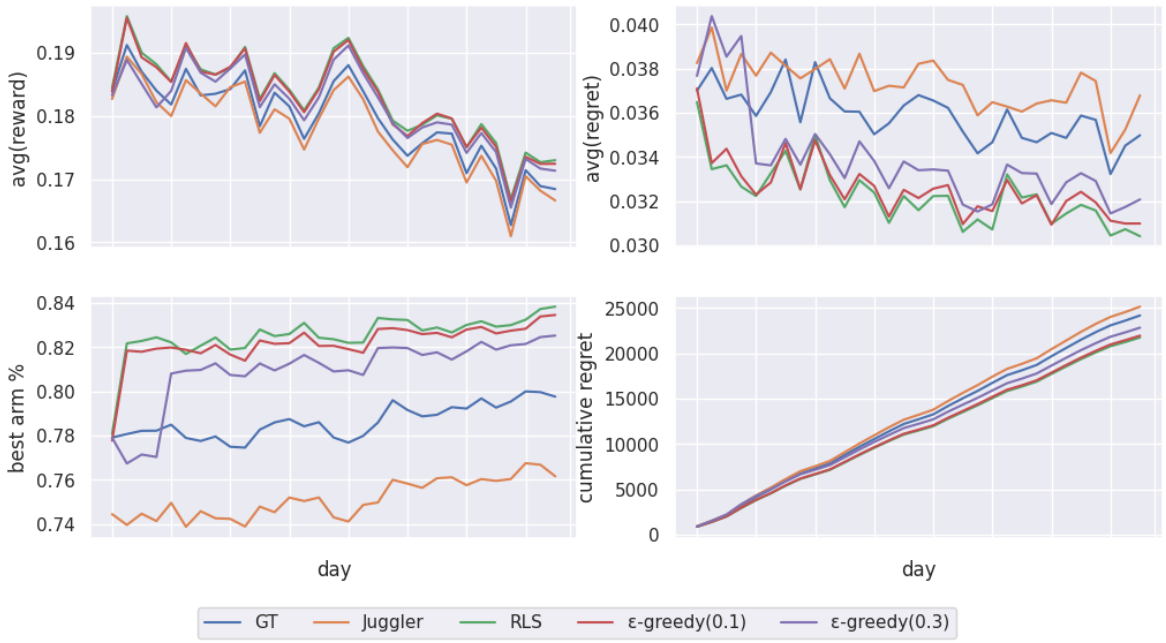

Figure 2: Multiple metrics per bandit over time.

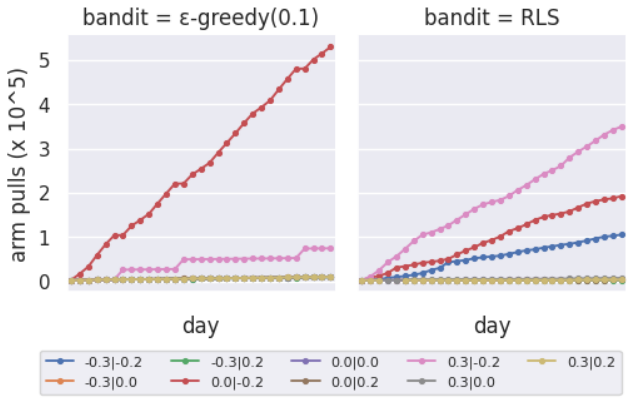

Figure 3: Arm pulls per bandit over time.

6. Conclusion and Future Work

In this paper, we presented a novel hybrid approach combining Meta-Learning with Multi-Armed Bandits for multi-stakeholder recommendations in online travel marketplaces. Our Juggler-MAB system demonstrated significant improvements over existing methods. Key contributions of our work include 1) an integration of meta-learning and contextual bandits for recommendation systems and 2) empirical evidence of the effectiveness of our approach in a large-scale, real-world setting. Based on our findings and the limitations identified, we propose the following directions for future research:

1. Online testing: Conduct A/B tests in a production environment to validate the performance of Juggler-MAB under real-world conditions and user behaviors
2. Dynamic arm space: Explore methods for dynamically adjusting the arm space of the MAB component based on observed performance and changing market conditions.
3. Fairness considerations: Incorporate explicit fairness constraints or objectives into the MAB formulation to ensure equitable treatment of different provider segments [21]
4. Long-term value optimization: Extend the approach to consider long-term user value, potentially using reinforcement learning techniques for sequential decision-making.

References

[1] H. Abdollahpouri, G. Adomavicius, R. Burke, I. Guy, D. Jannach, T. Kamishima, J. Krasnodebski, L. Pizzato, Multistakeholder recommendation: Survey and research directions, User Modeling and User-Adapted Interaction 30 (2020) 127–158.

[2] H. Abdollahpouri, R. Burke, Multi-stakeholder recommendation and its connection to multi-sided fairness, in: Workshop on Recommendation in Multi-stakeholder Environments (RMSE'19), in Conjunction with the 13th ACM Conference on Recommender Systems, RecSys'19, 2019.

[3] R. Mehrotra, B. Carterette, Recommendations in a marketplace, in: Proceedings of the 13th ACM Conference on Recommender Systems, 2019, pp. 580–581.

[4] D. Jannach, G. Adomavicius, Recommendations with a purpose, in: Proceedings of the 10th ACM Conference on Recommender Systems, ACM, 2016, pp. 7–10.

[5] T. Cunha, I. Partalas, P. Nguyen, Juggler: Multi- stakeholder ranking with meta-learning, in: Proceedings of the MORS workshop at the 15th ACM Conference on Recommender Systems, CEUR Workshop Proceedings, 2021.

[6] T. Cunha, C. Soares, A. C. de Carvalho, Metalearning and recommender systems: A literature review and empirical study on the algorithm selection problem for collaborative filtering, Information Sciences 423 (2018) 128–144.

[7] T. Cunha, C. Soares, A. C. de Carvalho, Cf4cf: Recommending collaborative filtering algorithms using collaborative filtering, in: RecSys 2018 - Proceedings of the 12th ACM Conference on Recommender Systems, 2018, p. 357–361.

[8] T. Lattimore, C. Szepesvári, Bandit algorithms, Cambridge University Press (2020).

[9] L. Li, W. Chu, J. Langford, R. E. Schapire, A contextual-bandit approach to personalized news article recommendation, in: Proceedings of the 19th international conference on World wide web, 2010, pp. 661–670.

[10] H. Wang, Q. Wu, H. Wang, Factorization bandits for interactive recommendation, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 31, 2017.

[11] F. Hutter, H. H. Hoos, K. Leyton-Brown, Sequential model-based optimization for general algorithm configuration, International conference on learning and intelligent optimization (2011) 507–523.

[12] S. Falkner, A. Klein, F. Hutter, Bohb: Robust and efficient hyperparameter optimization at scale, in: International Conference on Machine Learning, PMLR, 2018, pp. 1437–1446.

[13] M. Rodriguez, C. Posse, E. Zhang, Multiple objective optimization in recommender systems, in: Proceedings of the sixth ACM conference on Recommender systems, 2012, pp. 11–18.

[14] P. Nguyen, J. Dines, J. Krasnodebski, A multi-objective learning to re-rank approach to optimize online marketplaces for multiple stakeholders, arXiv preprint arXiv:1708.00651 (2017).

[15] Ö. Sürer, R. Burke, E. C. Malthouse, Multistakeholder recommendation with provider constraints, in: Proceedings of the 12th ACM Conference on Recommender Systems, ACM, 2018, pp. 54–62.

[16] W. Black, E. Ilhan, A. Marchini, V. Markeviciute, Adaptex: A self-service contextual bandit platform, in: Proceedings of the 17th ACM Conference on Recommender Systems, 2023, pp. 839–842.

[17] E. Ie, V. Jain, J. Wang, S. Navrekar, R. Agarwal, R. Wu, H.-T. Gao, M. Chandra, C. Boutilier, Recsim: A configurable simulation platform for recommender systems, arXiv preprint arXiv:1909.04847 (2019).

[18] J. Demšar, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res. 7 (2006) 1–30.

[19] M. Dudík, J. Langford, L. Li, Doubly robust policy evaluation and learning, ICML'11, Omnipress, Madison, WI, USA, 2011, p. 1097–1104.

[20] A. Swaminathan, T. Joachims, The self-normalized estimator for counterfactual learning, advances in neural information processing systems 28 (2015).

[21] R. Burke, N. Sonboli, A. Ordoñez-Gauger, Balanced neighborhoods for multi-sided fairness in recommendation, in: Conference on Fairness, Accountability and Transparency, PMLR, 2018, pp. 202–214.