KG-TS: Knowledge Graph-driven Thompson Sampling for Online Recommendation

Cairong Yan, Hualu Xu, Yanting Zhang, Zijian Wang, and Xuan Shao(✉)

School of Computer Science and Technology, Donghua University, Shanghai, China
{cryan,ytzhang,wang.zijian,shx}@dhu.edu.cn,hualuxu@mail.dhu.edu.cn

Abstract. To address the challenges of sparse data and high dynamics in Contextual Multi-Armed Bandits (CMAB) models for online recommendation, this study introduces a novel Knowledge Graph-driven Thompson Sampling (KG-TS) algorithm within the CMAB framework. This algorithm innovatively constructs a dynamic Knowledge Graph (KG) that links user characteristics to item attributes, converting sequential decision-making into graph structures to explore data relationships and enhance contextual understanding. Additionally, a time-varying reward mechanism dynamically adjusts the edge weights of the KG, enabling more adaptive and timely personalization in recommendations. Theoretical analysis confirms that KG-TS achieves sublinear cumulative regret growth, demonstrating its efficacy in maximizing long-term benefits. Extensive experiments conducted on two public datasets show that our algorithm outperforms existing bandit algorithms by more than doubling the F1 score and reducing the regret value by over 10%, thus affirming its superior effectiveness in the online recommendation domain.

Keywords: knowledge graph · thompson sampling · non-stationary environment · online recommendation.

1 Introduction

Online recommendation plays a pivotal role in domains such as information retrieval [16], e-commerce [14, 12], and clinical medicine [7]. Their primary goal is to enhance user experience and augment platform revenue by offering personalized feedback-driven services. A fundamental challenge in this area is to achieve a balance between exploring new items and mitigating associated risks. To address this challenge, particularly in improving predictions and managing the exploration-exploitation trade-offs, Contextual Multi-Armed Bandits (CMAB) are extensively employed. As a variant of reinforcement learning within the multi-armed bandits framework, CMAB is instrumental in devising effective strategies. In this methodology, each potential item is conceptualized as an "arm" with its rewards governed by probability distributions. Bandit algorithms strategically alternate between arms, navigating the interplay of exploration and exploitation based on user feedback, thereby progressively refining the recommendation.

To address challenges such as user behavior poor and cold start issues in recommender systems, certain bandit algorithms integrate KG [15]. These graphs

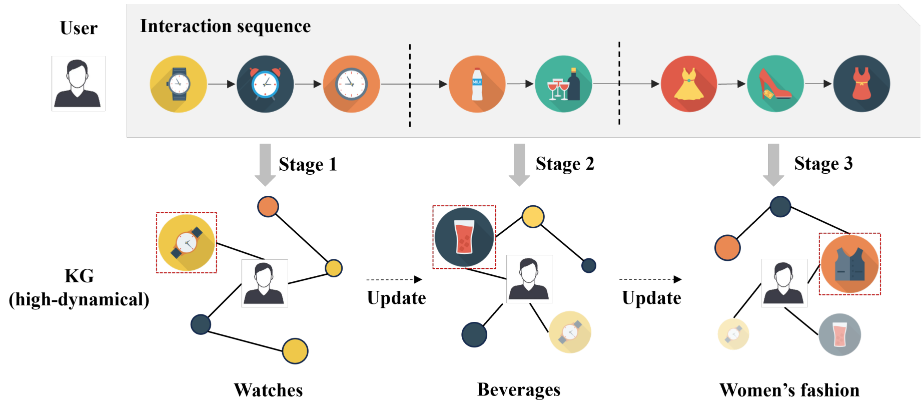

Fig. 1. An example of constructing a KG from user behavior to aid recommendation.

offer structured insights into users, items, and their interrelations, thereby furnishing the essential contextual information required for delivering precise and personalized recommendation. However, the associations between users and items in these bandit algorithms are predetermined, and face non-stationary scenarios where user preferences are subject to change over time. Fig. 1 illustrates this concept through a sparse sequence of user interactions, highlighting how user preferences can evolve: transitioning from an interest in watches to beverages, and eventually to women's fashion, while previous interests diminish.

In response to the above challenges, we propose a novel Knowledge Graph-driven Thompson Sampling (KG-TS) algorithm tailored to address the challenges of data sparsity and cold start issues in online recommender systems. Our contributions are summarized as follows:

1) The KG-TS algorithm is a pioneering work to integrate a KG as supplementary information within the CMAB framework for online recommendation. This approach distinctively combines user characteristics and item attributes through the graph's structure. A rigorous theoretical analysis supports the algorithm's effectiveness, achieving sublinear growth in cumulative regret.

2) We develop a flexible, Time-Varying Reward Mechanism (TV-RM) capable of adapting to the changing preferences of users. This mechanism employs a decay approach to progressively reduce the impact of previous behaviors, while simultaneously accounting for possible recurring needs derived from multi-faceted implicit feedback. Such a dynamic system ensures that the rewards are quickly and efficiently reflected in the KG's edge weights, allowing for effective adaptation to non-stationary conditions.

3) Empirical evaluations using two real-world datasets reveal that our algorithm consistently surpasses contemporary benchmarks in online recommendation scenarios. The performance is particularly pronounced in overcoming the pivotal challenges associated with data scarcity and fluctuating environments.

2 Methodology

2.1 Problem Formulation

We model the sequential recommendation task in e-commerce as a CMAB problem. We assume that there are M users $U = \{u_1, u_2, ..., u_M\}$ and N arms $I = \{i_1, i_2, ..., i_N\}$. Here one arm corresponds to one item and is associated with a d-dimensional feature vector x_t. With the above definition, the reward r_t is generated by a function $x_t^T \mu_t$, where $\mu_t \in R^d$ is a stationary but unknown parameter sampling from Gaussian distribution $N(\hat{\mu}_{t-1}, V_{t-1}^{-1})$:

$$\hat{\mu}_{t-1} = V_{t-1}^{-1} \textstyle\sum_{\tau=1}^{t-1} x_\tau r_\tau, \ V_{t-1} = I_d + \sum_{\tau=1}^{t-1} x_\tau x_\tau^T, \tag{1}$$

where I_d denotes a d-dimensional unit vector. Combining the prior $N(\hat{\mu}_{t-1}, V_{t-1}^{-1})$ and likelihood function, due to the conjugate property of the Gaussian distribution [3], the posterior can be rewritten as $N(\hat{\mu}_t, V_t^{-1})$. The agent always chooses to pull the arm $a^* = argmax_{a \in A} x_t^T \mu_t$ with the highest expected reward and recommends the item. Finally, the actual reward is calculated based on user feedback to update the KG and recalculate the reward distribution of the associated arms, before proceeding to the next round of recommendations.

Additionally, we define regret $R(t)$ as the difference between the expected reward $x_t^{*T} \mu_t^*$ of the best arm a^* and the actual reward $x_t^T \mu_t$ of the pulled arm a_t. The agent aims to minimize the cumulative regret $R(\mathcal{T})$ over time $\mathcal{T}$:

$$R(\mathcal{T}) = \textstyle\sum_{t=1}^{\mathcal{T}} R\left(t\right) = \sum_{t=1}^{\mathcal{T}} \left(x_t^{*T} \mu_t^* - x_t^T \mu_t\right) . \tag{2}$$

2.2 Graph-based Contextual Representation

To address the recommendation problem in sparse scenarios, this section proposes a knowledge graph-enhanced CMAB model, as shown in Fig. 2.

1) Construction and updating of KG. In e-commerce, historical user feedback is often sparse, limiting accurate context extraction. However, recommendations depend heavily on contextual information, highlighting the importance of leveraging KGs to enhance context in CMAB research.

While traditional KGs [10] focus on capturing direct interactions between users and items, our objective is to extract deeper connections. We initially define a four-dimensional matrix $E \in \mathbb{R}^{U \times I \times P \times Q}$, where U represents the set of user characteristics f_u(e.g., age, gender, and region), I represents the set of item attributes f_i(e.g., category, brand, and price), and P and Q denote values $f_{u,p}$ and $f_{i,q}$ corresponding to user characteristics and item attributes, respectively.

We employ E as the foundational structure for constructing the KG, each user characteristic or item attribute value is treated as an entity. The connections and weights between entity nodes are determined based on the values $\bar{e}_{u,p}^{i,q}$ in the matrix E, thereby generating knowledge triples $(f_{u,p}, \bar{e}_{u,p}^{i,q}, f_{i,q})$. First, $e_{u,p}^{i,q}(t)$ is defined to represent the change in the additional relationship between $f_{u,p}$ and $f_{i,q}$ caused by a user's behavioral feedback at a given time. This value is

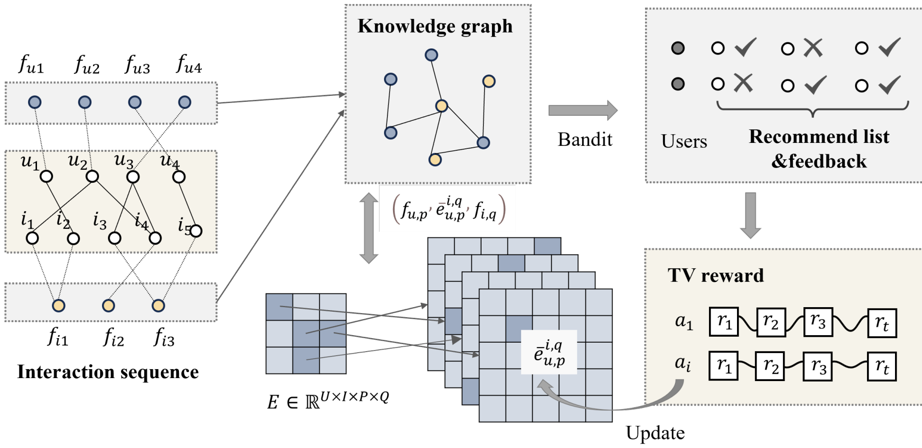

Fig. 2. Knowledge graph-enhanced CMAB recommendation process.

equivalent to the reward r_t generated in a specific round of recommendation due to the selection of a particular arm by the multi-armed bandit. For any user u and item/arm i, the formula is defined as follows:

$$e^{i,q}_{u,p}(t) = \mathbb{1}(I_u = i)g(k_i) \ , \tag{3}$$

where $\mathbb{1}(I_u = i)$ is used to indicate whether the user u pulls the arm i, with a return value of 0 or 1. And $g(k_i)$ is used to updating the KG, represents a function for calculating feedback scores, which will be detailed in Section 2.3. Then we define $\bar{e}^{i,q}_{u,p}$ as the proximity between $f_{u,p}$ and $f_{i,q}$, representing the edge weights between these entities:

$$\bar{e}^{i,q}_{u,p} = \textstyle\sum_{\tau=1}^{t} e^{i,q}_{u,p}(\tau) \ . \tag{4}$$

2) Knowledge combination reward distribution. The reward distribution, a key component of the CMAB, governs the probability of selecting actions in an explore/exploit manner, significantly impacting the algorithm's recommendation effectiveness. Based on the KG structure and update mechanisms, the KG edges are integrated as reward distribution parameters in the CMAB.

Assuming each arm corresponds to an item, the distribution $N(\hat{\mu}_t, V_t^{-1})$ for user u in the contextual Thompson Sampling algorithm is enhanced by incorporating the edge relationships $e^{i,q}_{u,p}(\tau)$ from the KG as contextual information. The mean of distribution can be expressed as:

$$\hat{\mu}_t = V_t^{-1} \textstyle\sum_{\tau=1}^{t-1} \left(x_\tau \sum_{u,i=1}^{U,I} \eta_{u,i} e^{i,q}_{u,p}(\tau) \right) \ , \tag{5}$$

where $\eta_{u,i}$ denotes the focus of interest for the user-item pair. For example, students might prioritize price, while workers may focus more on quality, leading to different attention levels toward item attributes for different users. Here, we assign them the same weight, i.e., $\eta_{u,i}$ is always set to 1. We refer to the method of aggregating multiple sets of knowledge triplets based on user features into a single reward distribution as "Knowledge Combination Recommendation".

2.3 Time-varying Reward Mechanism

User interests evolve over time, leading to a phenomenon known as "drift" in rewards related to arms. In non-stationary environments, the KG needs continuous updates. therefore, we propose a Time-Varying Reward Mechanism (TV-RM):

1) Periodic forgetting strategy. In user-item interactions, users' interest in items tends to diminish over time, especially when there is no subsequent interaction. However, once the user interacts again, interest is regained and then gradually declines based on the new memory. We map the fluctuating user interest to the average reward of the arms, treating the degeneration of interest between interactions as a cyclical process with periodic decay over time.

Initially, we employ a time-dependent *sigmoid* function. The purpose of this function is to diminish the influence of past rewards. This reduction is crucial to prevent an unbounded escalation in reward values and to mitigate the natural decline of user attention over time. Subsequently, we define n_t to denote the number of times each arm a has not been pulled recently: For the arm pulled in the current round, n_t is incremented by 1 to indicate an increase in the pull count, while for the arms not pulled, n_t is reset to a minimal value ε:

$$n_t = \varepsilon + \mathbb{1}(I_t \neq a)\textstyle\sum_{\tau=j}^{t} \mathbb{1}(I_\tau \neq a) \ , \tag{6}$$

where ε is set to 0.01, j represents the moment when the arm was last pulled, and $\mathbb{1}(I_t \neq a)$ indicates whether arm a was pulled. Let $f_t = \sum_{\tau=1}^{t-1} x_\tau r_\tau$ denote a part of the composition of $\hat{\mu}_t$. Then, we get $f_t = f_{t-1} \cdot sigmoid({n_t}^{-1}) + x_t r_t$.

Based on the above, we assign a minimal value to n_t for recently activated arms, indicating peak user interest. Conversely, for the unchosen arms, the reward gradually decrease. n_t serves as an indirect indicator of periodic reward fluctuation, capturing the dynamics of user interest and engagement.

2) Adaptive reward weight setting strategy. We categorize user interactions into strong and weak, forming the basis for reward. Strong interaction arms, denoted as A^s, include actions like multiple clicks, favorites, add-to-cart, or purchases, indicating user interest and triggering positive feedback (value 1). Conversely, weak interaction arms A^w, are characterized by no interaction or a single click, resulting in negative feedback (value 0). The symbol $s_t = \mathbb{1}(a \in A^s)$ represents this implicit feedback. We allocate a reward weight $\varDelta\bar{w}_t$ to the expected arm a^{opt}, such that its reward mean μ_{t+1}^{opt} maintained will have the potential to outperform the arm a^* with the current maximum reward mean μ_t^*, thus facilitating accurate recommendation in the next round. Accordingly, we let $\varDelta w_t = (\mu_t^* - \mu_t^{opt}) + \xi$. To avoid storing a large volume of historical operations and to accommodate slow changes, we update the weights in a fully recursive manner to correct the latest reward weights $\varDelta\bar{w}_t$:

$$\varDelta\bar{w}_t = ((t-1)\varDelta\bar{w}_{t-1} + \varDelta w_t) \cdot t^{-1} \ , \tag{7}$$

Based on the above definition of s_t, we obtain a weighted time-varying bandit algorithm with a general nonlinear reward $\mathbb{r}_t = s_t \Delta\bar{w}_t$. In this manner, f_t is updated as follows: $f_t = f_{t-1} \cdot sigmoid\left({n_t}^{-1}\right) + x_t \mathbb{r}_t$.

2.4 KG-TS Algorithm

CMAB achieves lower cumulative regret [17]. Building on this, we propose the KG-TS algorithm (Algorithm 1), integrating Thompson Sampling (TS) with a Knowledge Graph (KG).

Algorithm 1: KG-TS algorithm

Input: Arm set A, contextual vector x_t, Relational matrix $E \in \mathbb{R}^{U\times I\times P\times Q}$
Output: Recommendation list L
Init: $f_0 = 0_d$, $V_0 = I_d$, $\hat{\mu}_0 = 0_d$, $n_0 = \varepsilon$, $\eta_{u,i} = 1$
for $t = 1, 2, 3, ..., \mathcal{T}$ **do**
  Sampling μ_t from distribution $N(\hat{\mu}_{t-1},\ V_{t-1}^{-1})$
  Choose arm $a_t^* = argmax_{a\in A} x_t^T \mu_t$
  Get recommendation list $L = \{i_1, i_2, \ldots, i_K\}$ according to a_t^*
  Observe payoff $\mathbb{r}_t$:
    $s_t = \mathbb{1}(a \in A^s)$
    $\Delta w_t = (\mu_t^* - \mu_t^{opt}) + \xi$
    $\Delta\bar{w}_t = ((t-1)\Delta\bar{w}_{t-1} + \Delta w_t)\, t^{-1}$
    $\mathbb{r}_t = s_t \cdot \Delta\bar{w}_t$
  for $a \in A$ **do**
    if $\mathbb{1}\,(I_t \neq a)$ **then**
      $n_t = n_{t-1} + 1$
    else
      $n_t = \varepsilon$
  Update the knowledge graph E:
    $g(k_i) = \mathbb{r}_t$
    $e_{u,p}^{i,q}(t) = \mathbb{1}(I_u = i) g(k_i)$
    $\bar{e}_{u,p}^{i,q}(t) = \sum_{\tau=1}^{t} e_{u,p}^{i,q}(\tau)$
  Update V_t, f_t and $\hat{\mu}_t$ as follows:
    $f_t = f_{t-1} \cdot sigmod({n_t}^{-1}) + x_t \sum_{u,i=1}^{U,I} \eta_{u,i} e_{u,p}^{i,q}(\tau)$
    $V_t = V_{t-1} + x_t x_t^T$
    $\hat{\mu}_t = V_t^{-1} f_t$

Regret Analysis We will prove that the KG-TS algorithm has sublinear properties. First, we give the following definitions to help the subsequent proof:

Definition 1. *Super-martingale[1].* $X_t = \mathbb{I}\left(N_{x_t^*} > 0\right) x_t^{*T} \mu_t^* - \mathbb{I}\left(N_{x_t} > 0\right) x_t^T \mu_t$, $Y_t = \sum_{\tau=1}^{t} X_\tau$, *where* Y_t *denotes the super-martingale process.*

Definition 2. *Confidence Ellipsoid [5]. Define* $\beta_t(\delta) = R\sqrt{dln\frac{1+tL/\lambda}{\delta}} + \sqrt{\lambda L}$, *we can get* $\parallel \hat{\mu}_t - \mu_t \parallel_{V_t} \leq \beta_t(\delta)$ *with probability* $1-\delta$ *for any* $0 \leq \delta \leq 1$.

Step 1. Assuming that $0 \leq \| x_t \| \leq 1$, $0 \leq \| \mu_t \| \leq 1$, so that the inclusion of knowledge-enhanced contextual information will not affect the original proof of regret. The inequality $0 \leq x_t^T \mu_t \leq 1$ is true, so $X_t \leq 1$. According to the Azuma–Hoeffding inequality, there is a probability of $1-\delta$ that $Y_{\mathcal{T}} \leq 1\sqrt{2\mathcal{T} \ln \frac{2}{\delta}}$.

Step 2. According to the Cauchy–Schwarz inequality and the Deduction of Confidence Ellipsoid in Definition. 2, for any vector x_t, we can get:

$$x_t^T \left(\mu_t - \hat{\mu}_t\right) = x_t^T V_t^{-\frac{1}{2}} V_t^{\frac{1}{2}} \left(\mu_t - \hat{\mu}_t\right) \leq \| x_t \|_{V_t^{-1}} \beta_t\left(\delta\right) \ . \tag{8}$$

Step 3. For each round, the regret value can be calculated by:

$$R\left(t\right) = X_t + \left(x_t^{*T} \mu_t - x_t^{*T} \hat{\mu}_t\right) - \left(x_t^T \mu_t - x_t^T \hat{\mu}_t\right) \ , \tag{9}$$

$$R\left(\mathcal{T}\right) \leq Y_{\mathcal{T}} + \sum_{t=1}^{\mathcal{T}} \left(\| x_t^* \|_{V_t^{-1}} + \| x_t \|_{V_t^{-1}}\right) \beta_t\left(\delta\right) \ . \tag{10}$$

Step 4. According to the previous definition of $Y_{\mathcal{T}}$, $\| x_t \|_{V_t^{-1}}$ and $\beta_t\left(\delta\right)$, let $\delta = \frac{\gamma}{6(\mathcal{T}+1)^2}$ replace the parameter, we can get:

$$Y_{\mathcal{T}} \leq 1\sqrt{2\mathcal{T} \ln \frac{2}{\delta}} = \sqrt{2\mathcal{T} \ln \frac{12\left(\mathcal{T}+1\right)^2}{\gamma}} = O(\sqrt{\mathcal{T} \ln \frac{\mathcal{T}^2}{\gamma}}) \ , \tag{11}$$

$$\sum_{t=1}^{\mathcal{T}} \| x_t \|_{V_t^{-1}} = \sqrt{x_t V_t^{-1} x_t^T} = O(\sqrt{d\mathcal{T}}) \ , \tag{12}$$

$$\beta_t\left(\delta\right) = R\sqrt{d \ln \frac{1 + tL/\lambda}{\delta}} + \sqrt{\lambda L} = O(\sqrt{d \ln \mathcal{T}^3/\gamma}). \tag{13}$$

Step 5. Finally, we calculate:

$$\begin{aligned} R\left(\mathcal{T}\right) &= \sum_{t=1}^{\mathcal{T}} \left(x_t^{*T} \mu_t^* - x_t^T \mu_t\right) \leq Y_{\mathcal{T}} + \sum_{t=1}^{\mathcal{T}} \left(\| x_t^* \|_{V_t^{-1}} + \| x_t \|_{V_t^{-1}}\right) \beta_t\left(\delta\right) \\ &= O\left(\sqrt{\mathcal{T} \ln \frac{\mathcal{T}^2}{\gamma}} + \sqrt{d\mathcal{T}} \cdot \sqrt{\frac{d \ln \mathcal{T}^3}{\gamma}}\right) = O\left(d\sqrt{\mathcal{T} \ln \frac{\mathcal{T}}{\gamma}}\right), \end{aligned} \tag{14}$$

Here we can rewrite $O\left(d\sqrt{\mathcal{T} \ln \frac{\mathcal{T}}{\gamma}}\right)$ in the form of $\widetilde{O}(d\sqrt{\mathcal{T}})$ to show that the cumulative regret exhibiting sublinear property.

3 Experiment and Evaluation

We utilized two real-world datasets: IJCAI-2015 and UBD, which contain extensive user behavioral data from e-commerce platforms. For the evaluation, we used four metrics: precision, recall, F1, and cumulative regret. Finally, We compared KG-TS with seven baselines: ε-greedy [4], UCB [2], LinUCB [6], TS [13], Exp3.S [9], VarUCB [11], and SW-TS [8]. The experiments were designed to answer the following research questions: **RQ1**: Can the proposed KG-TS outperform classic and state-of-the-art bandit algorithms in recommendation tasks? **RQ2**: How does KG-TS mitigate the data sparsity problem? **RQ3**: How do the different components affect the performance of KG-TS?

Table 1. Performance comparison on online recommendation between the baselines and our model (all the values in the table are percentage numbers with % omitted).

Datasets	IJCAI-2015				UBD			
Metrics	Precision↑	Recall↑	F1↑	Regret↓	Precision↑	Recall↑	F1↑	Regret↓
ε-greedy	0.07	0.02	0.03	90348	0.05	0.02	0.03	91,714
UCB	0.18	0.11	0.13	91789	0.09	0.06	0.07	92,049
LinUCB	0.41	0.23	0.29	83104	0.21	0.16	0.18	88,948
TS	0.33	0.26	0.29	84521	0.17	0.10	0.12	88,236
Exp3.S	0.27	0.21	0.23	85207	0.19	0.07	0.10	89,971
VarUCB	1.69	1.29	1.46	73358	0.93	0.71	0.80	76,423*
SW-TS	2.76*	2.14*	2.41*	60073*	1.00*	0.98*	0.98*	76,649
KG-TS (Ours)	**5.88**	**4.11**	**4.83**	**52808**	**2.20**	**1.89**	**2.03**	**68,411**
Impr.	+3.12%	+1.97%	+2.42%	+12.09%	+1.20%	+0.91%	+1.05%	+10.48%

3.1 Performance Comparison (RQ1)

Table 1 provides a detailed presentation of the performance of all baselines in predicting target user interaction behaviors and exploration-exploitation capabilities on the IJCAI-2015 and UBD datasets. Through a comprehensive analysis, we can draw several key conclusions:

1) Classical bandit algorithms, specifically ε-greedy, UCB, and TS, exhibit relatively weaker performance. 2) LinUCB and varUCB leverage user and item features to estimate rewards and confidence intervals, significantly outperform UCB in terms of recommendation effectiveness and regret convergence. 3) KG-TS outperforms three time-varying bandit algorithms (Exp3.S, VarUCB, and SW-TS) in addressing non-stationary problem.

3.2 Solution to Data Sparsity Problem (RQ2)

As depicted in Tables 2, in terms of F1, the methods leveraging the KG significantly outperform traditional approaches that do not capture KG and contextual information, particularly in data-sparse scenarios.

Even on the two datasets, when the data volume is reduced to only one-tenth of the original scale, the KG-TS algorithm excels, with F1 scores decreasing by only 10.97% and 18.23%, respectively. In contrast, the suboptimal LinUCB algorithm experiences a reduction of 37.93% and 38.89%. Although Var-UCB exhibits moderate overall performance, considering both feature information and changes in user interests, its performance only diminishes by 28.08% and 22.50%.

3.3 Ablation Study (RQ3)

To evaluate the contributions of different model components, this experiment analyzed three scenarios: excluding all contextual information (w/o KG+TV), removing the KG module (w/o KG), and eliminating the time-varying reward mechanism (w/o TV). The results, shown in Fig 3.

Table 2. F1 of algorithms on the IJCAI-2015 and UBD dataset with reduced user interaction records (all the values in the table are percentage numbers with % omitted).

IJCAI-2015	100k	90k	80k	70k	60k	50k	40k	30k	20k	10k	Reduce↓
ε-greedy	0.03	0.03	0.03	0.03	0.03	0.02	0.02	0.02	0.01	0.01	66.67%
UCB	0.13	0.13	0.13	0.12	0.12	0.10	0.09	0.06	0.05	0.05	61.54%
LinUCB	0.29	0.29	0.28	0.27	0.27	0.25	0.20	0.19	0.18	0.18	37.93%
TS	0.29	0.28	0.27	0.25	0.25	0.21	0.19	0.15	0.13	0.12	58.62%
Exp3.S	0.23	0.21	0.20	0.20	0.19	0.17	0.16	0.15	0.15	0.14	39.13%
Var-UCB*	1.46	1.43	1.40	1.38	1.33	1.29	1.22	1.20	1.14	1.05	28.08%*
SW-TS	2.41	2.40	2.29	2.11	2.03	1.95	1.69	1.43	1.32	1.17	51.45%
KG-TS(Ours)	**4.83**	**4.82**	**4.79**	**4.71**	**4.62**	**4.50**	**4.46**	**4.41**	**4.37**	**4.30**	**10.97%**
UBD	100k	90k	80k	70k	60k	50k	40k	30k	20k	10k	Reduce↓
ε-greedy	0.03	0.03	0.02	0.02	0.02	0.02	0.02	0.02	0.01	0.01	66.67%
UCB	0.07	0.07	0.07	0.07	0.06	0.06	0.04	0.05	0.04	0.03	57.14%
LinUCB	0.18	0.17	0.17	0.17	0.16	0.15	0.13	0.12	0.12	0.11	38.89%
TS	0.12	0.12	0.12	0.12	0.10	0.10	0.09	0.08	0.07	0.07	41.67%
Exp3.S	0.10	0.10	0.09	0.09	0.09	0.08	0.08	0.07	0.06	0.04	60.00%
Var-UCB*	0.80	0.81	0.79	0.74	0.71	0.69	0.68	0.64	0.63	0.62	22.50%*
SW-TS	0.98	0.96	0.91	0.85	0.81	0.77	0.76	0.62	0.58	0.45	54.08%
KG-TS(Ours)	**2.03**	**2.00**	**1.92**	**1.86**	**1.82**	**1.74**	**1.71**	**1.69**	**1.67**	**1.66**	**18.23%**

4 Conclusion

In this paper, we propose KG-TS for online recommendation. This algorithm adopts a dual decision-making approach: 1) Leveraging a dynamic KG to provide contextual information for bandit decision-making, with an innovative schema that models features as entities and incorporates a time-varying reward mechanism for KG updates; 2) Utilizing knowledge triples to allocate rewards, enabling precise estimation of expected returns. Both theoretical and experimental results validate the competitive advantage of KG-TS. Future work will explore fine-grained KG construction techniques to capture richer contextual information.

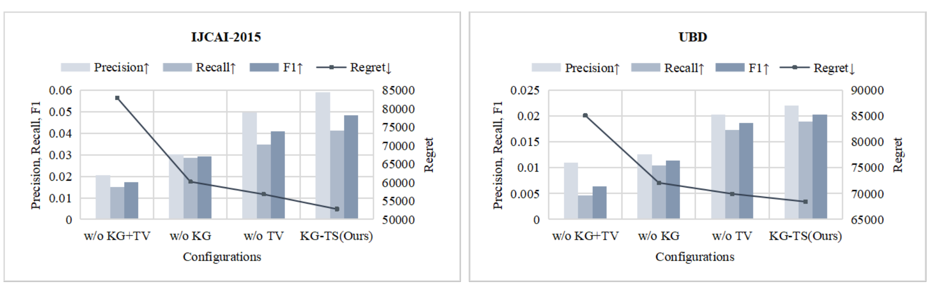

Fig. 3. Ablation study of KG-TS.

Acknowledgements. This work is supported by the National Natural Science Foundation of China under Grant 62477006 and 62206046.

References

1. Agrawal, S., Goyal, N.: Thompson sampling for contextual bandits with linear payoffs. In: Proceedings of the International Conference on Machine Learning. pp. 127–135 (2013)
2. Auer, P., Cesa-Bianchi, N., Fischer, P.: Finite-time analysis of the multiarmed bandit problem. Machine Learning **47**, 235–256 (2002)
3. Hong, J., Kveton, B., Zaheer, M., Ghavamzadeh, M., Boutilier, C.: Thompson sampling with a mixture prior. In: Proceedings of the International Conference on Artificial Intelligence and Statistics. pp. 7565–7586 (2022)
4. Kuleshov, V., Precup, D.: Algorithms for multi-armed bandit problems. arXiv preprint arXiv:1402.6028 (2014)
5. Li, C., Wang, H.: Asynchronous upper confidence bound algorithms for federated linear bandits. In: Proceedings of the International Conference on Artificial Intelligence and Statistics. pp. 6529–6553 (2022)
6. Li, L., Chu, W., Langford, J., Schapire, R.E.: A contextual-bandit approach to personalized news article recommendation. In: Proceedings of the International Conference on World Wide Web. pp. 661–670 (2010)
7. Tomkins, S., Liao, P., Klasnja, P., Murphy, S.: Intelligentpooling: Practical thompson sampling for mhealth. Machine Learning **110**(9), 2685–2727 (2021)
8. Trovo, F., Paladino, S., Restelli, M., Gatti, N.: Sliding-window thompson sampling for non-stationary settings. Artificial Intelligence Research **68**, 311–364 (2020)
9. Vakili, S., Zhao, Q., Zhou, Y.: Time-varying stochastic multi-armed bandit problems. In: Proceedings of the Asilomar Conference on Signals, Systems and Computers. pp. 2103–2107 (2014)
10. Wang, X., He, X., Cao, Y., Liu, M., Chua, T.S.: Kgat: Knowledge graph attention network for recommendation. In: Proceedings of the ACM SIGKDD International Conference. pp. 950–958 (2019)
11. Xu, L., Jiang, C., Qian, Y., Zhao, Y., Li, J., Ren, Y.: Dynamic privacy pricing: A multi-armed bandit approach with time-variant rewards. IEEE Transactions on Information Forensics and Security **12**(2), 271–285 (2016)
12. Yan, C., Han, H., Zhang, Y., Zhu, D., Wan, Y.: Dynamic clustering based contextual combinatorial multi-armed bandit for online recommendation. Knowledge-Based Systems **257**, 109927 (2022)
13. Yan, C., Xian, J., Wan, Y., Wang, P.: Modeling implicit feedback based on bandit learning for recommendation. Neurocomputing **447**, 244–256 (2021)
14. Yan, C., Xu, H., Han, H., Zhang, Y., Wang, Z.: Thompson sampling with time-varying reward for contextual bandits. In: Proceedings of the International Conference on Database Systems for Advanced Applications. pp. 54–63 (2023)
15. Yang, Y., Huang, C., Xia, L., Li, C.: Knowledge graph contrastive learning for recommendation. In: Proceedings of the International ACM SIGIR Conference. pp. 1434–1443 (2022)
16. Yue, Y., Joachims, T.: Interactively optimizing information retrieval systems as a dueling bandits problem. In: Proceedings of the Annual International Conference on Machine Learning. pp. 1201–1208 (2009)
17. Zhu, Z., Huang, L., Xu, H.: Self-accelerated thompson sampling with near-optimal regret upper bound. Neurocomputing **399**, 37–47 (2020)