Expert Systems With Applications 197 (2022) 116669

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Multi-Armed Bandits in Recommendation Systems: A survey of the state-of-the-art and future directions

Nícollas Silva [a,*], Heitor Werneck [b], Thiago Silva [b], Adriano C.M. Pereira [a], Leonardo Rocha [b]

[a] Universidade Federal de Minas Gerais, Belo Horizonte, Brazil
[b] Universidade Federal de São João del-Rei, São João del-Rei, Brazil

ARTICLE INFO

Keywords:
Recommender Systems
Multi-Armed Bandits
Systematic literature review
Recommendation challenges

ABSTRACT

Recommender Systems (RSs) have assumed a crucial role in several digital companies by directly affecting their key performance indicators. Nowadays, in this era of big data, the information available about users and items has been continually updated and the application of traditional batch learning paradigms has become more restricted. In this sense, the current efforts in the recommendation field have concerned about this online environment and modeled their systems as a Multi-Armed Bandit (MAB) problem. Nevertheless, there is not a consensus about the best practices to design, perform, and evaluate the MAB implementations in the recommendation field. Thus, this work performs a systematic literature review (SLR) to shed light on this new topic. By inspecting 1327 articles published from the last twenty years (2000–2020), this work: (1) consolidates an updated picture of the main research conducted in this area so far; (2) highlights the most used concepts and methods, their core characteristics, and main limitations; and (3) evaluates the applicability of MAB-based recommendation approaches in some traditional RSs' challenges, such as data sparsity, scalability, cold-start, and explainability. These discussions and analyzes also allow us to identify several gaps in the current literature, providing a strong guideline for future research.

1. Introduction

In the last three decades, the exponential growth of digital information on the Web has induced users to a stressful situation in which they do not know what to buy, listen to, or to watch. This problem is known in the literature as *information overload* and it has influenced several researchers to work on Recommendation Systems (RSs) to provide suggestions of items (e.g., movies, books, songs, etc.) and mitigate this problem (Shapira, Ricci, Kantor, & Rokach, 2011). Formally, RSs aim to estimate the user's preference or even a specific *rating* for the available items in order to provide recommendations that increase both the user's satisfaction and the system's profit (Pathak, Garfinkel, Gopal, Venkatesan, & Yin, 2010). Distinct algorithms have been proposed so far based on the main recommendation strategies, such as Collaborative Filtering (CF), Demographic Filtering (DF), Content-based (CB), and Knowledge-based (KB) (Bobadilla, Ortega, Hernando, & Gutiérrez, 2013; Jannach, Zanker, Felfernig, & Friedrich, 2010; Park, Kim, Choi, & Kim, 2012).

Current efforts have proposed to handle the online recommendation task with concepts from the Reinforcement Learning (RL) field by modeling it as a Multi-Armed Bandit (MAB) problem (Wang, Wu, & Wang, 2016; Wang, Wu and Wang, 2017; Wu, Wang, Gu, & Wang, 2016; Zhao, Zhang, & Wang, 2013). Traditionally, MAB is defined as a sequential decision model that has to continually choose an action a among a set of actions $\mathcal{A}$ – a.k.a. arms. The selection of action $a \in \mathcal{A}$ in a trial t brings out in a certain reward $R_t(a_t) \in \mathbb{R}$, which can be summarized as a real number. The main goal is to maximize the reward returned $\sum_{t=1}^{T} R_t(a_t)$ for T trials. In the recommendation domain, items available are usually modeled as the arms to be pulled. Selecting an arm is equivalent to recommending an item, and the reward is the user's response (e.g., clicks, acceptance, satisfaction, etc.). Similar to traditional RL scenarios, to achieve its goal, the bandit model should balance the **exploitation and exploration dilemma**. While exploitation just means pull arms with the highest rewards in the past, maximizing the system short-term reward, exploration is achieved by recommending other arms to improve the knowledge available about users and items to maximize the system long-term reward (Sanz-Cruzado, Castells, & López, 2019; Zhao et al., 2013).

The MAB problem has attracted a lot of attention from both industry and academy in the recommendation field. In the academy, it is possible to notice that more than 50% of all publications about this topic was only proposed in the last five years, as shown in Fig. 1. Similarly,

* Corresponding author.
E-mail addresses: ncsilvaa@dcc.ufmg.br (N. Silva), werneck@aluno.ufsj.edu.br (H. Werneck), thiagosilva@aluno.ufsj.edu.br (T. Silva), adrianoc@dcc.ufmg.br (A.C.M. Pereira), lcrocha@ufsj.edu.br (L. Rocha).

https://doi.org/10.1016/j.eswa.2022.116669
Received 25 October 2021; Received in revised form 21 January 2022; Accepted 8 February 2022
Available online 19 February 2022
0957-4174/© 2022 Published by Elsevier Ltd.

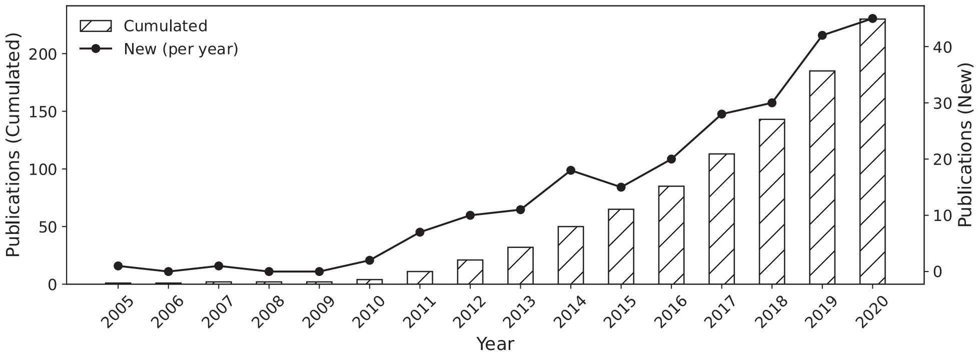


Fig. 1. Annual publications about MAB in the recommendation field identified through the Google Scholar tool according to our literature search method explained in next sections.

in the industry, recent talks of research leaders of Netflix, Pandora, and Spotify in the main conferences, such as ACM Recommender Systems (RecSys), ACM Conference on Research and Development in Information Retrieval (SIGIR), and Web Conference (WWW), have revealed the growing interest of companies on this topic to handle the online recommendation task, especially.

However, even with this growing number of new publications available, there is no work that aims to map the main advances in the area, explain the main concepts, and clarify the best practices. Therefore, this work performs a systematic literature review (SLR) about MAB in the recommendation field to achieve three main goals:

(1) Provide a summary overview of the most important research in this area;
(2) Highlight the most popular concepts and methods, their core characteristics, and their main limitations to provide future directions of the field to guide the next research questions;
(3) Discuss the applicability of MAB-based recommendation approaches in some traditional RSs' challenges, such as the data sparsity, scalability, cold-start, privacy, and explainability.

Searching all conferences available in the Google Scholar from 2000 to 2020, our SLR identified 1327 articles based on three main strings designed according to our goals and research questions. Then, we conducted two main reading steps to filter and identify the most relevant studies. While the first step performs a short reading by analyzing titles, year, conference, and abstract, the second one performs a more complete analysis by reading the introduction, experimentation, and conclusion of each work. In the first reading, only 408 papers (30.75%) were selected for the second step. Then, in the second stage, other 178 papers were rejected and only 230 papers were selected as relevant studies about MAB in the recommendation field.[1] These works were deeply studied to achieve our three main goals. They were read to improve our knowledge but only those with an experimental setup were used to fill a data extraction form designed to catch the current practices in the literature.

In general, the application of our SLR provides distinct contributions for academia and industry. In this paper, we highlight: (1) the main conferences where MAB studies have been published; (2) the most usual scenarios simulated by the publications; (3) the datasets applied for these studies; (4) the main algorithms usually applied to address the MAB problem; and (5) the main advances by combining traditional bandit algorithms with concepts of recommendation systems. Our work also identifies several gaps in the current literature and proposes relevant future directions. For instance, we noticed the absence of a strict evaluation criteria that reflect the traditional RS goals. Relevant metrics usually related to user satisfaction or engagement have been neglected by only applying the traditional evaluation criteria of learning algorithms based on rewards (or regrets). Moreover, it is not clear how the most relevant challenges of the recommendation field can affect bandit algorithms. We discussed common problems that still are trends in the field, like sparsity, scalability, cold-start, privacy, and explainability, by pointing out the future directions for research in these topics.

The remainder of this paper is organized as follows. First, Section 2 highlights background concepts about the traditional MAB problem. Then, Section 3 presents the SLR process by showing each step performed by our inspection. Section 4 organizes the main discussion of our paper by highlighting the works developed so far, the current evaluation criteria, and how MAB has faced the current challenges of the field. These discussions allow us to point out the main future directions in Section 5. Finally, Section 6 presents our main conclusions.

2. Background concepts

The Multi-Armed Bandit (MAB) problem, sometimes called the K-armed bandit problem (Zhao, Xia, Tang and Yin, 2019), is a classic problem in which a fixed limited set of resources (arms) must be selected between competing choices to maximize their expected gain (reward). The name 'bandit' comes from imagining a gambler at a row of slot machines in a casino, who has to improve his/her profit by maximizing the sum of rewards earned through a sequence of lever pulls. Basically, at each trial, the gambler has to decide which machines to play, how many times to play each machine and in which order to play them, and whether to continue with the current machine or try a different machine. However, each machine provides a random reward from a probability distribution specific to that machine and the best way to solve this problem is handling a crucial dilemma, deciding for the *exploitation* of the machine that has the highest expected payoff and the *exploration* of other machines to get more information about the expected payoffs.

Formally, the MAB is a sequential decision model represented by a 3-tuple $\langle \mathcal{A}, \mathcal{R}, \mathcal{Q} \rangle$, where an agent has to continually choose an action $a \in \mathcal{A}$ for $T \in \mathbb{N}$ trials among a set of actions $\mathcal{A}$ ($|\mathcal{A}| = K$) in order to maximize the cumulative reward $\sum_{t=1}^{T} r_t$. Here, $r_t = \mathcal{R}_t(a_t)$ is the reward achieved when an action a is performed. An algorithm for MAB performs an action a at each trial t according to an action selection policy π. This policy is driven to follow a probability distribution, usually called the value function $\mathcal{Q}$, over each possible action a. The function $\mathcal{Q}$ is the main responsible to select (or not) an action a because it measures the expected reward – $\mathcal{Q}_t(a) = \mathbb{E}[r_t|a]$. Fig. 2 also formalizes this MAB definition, showing the agent and the environment interacting continually in a sequence of discrete time steps t. At each time step t, the agent samples an arm a_t and receives a reward r_t. The history of actions and rewards guides the agent selection for the time $t+1$ and the other trials.

In a similar definition, it is possible to redefine the MAB main goal to minimize the regret associated with each action a chosen, as

[1] The list with all works selected is available at: https://drive.google.com/file/d/1eDAZTCgFmToqy5iW11if_-6v7J0tJS3z/view?usp=sharing.

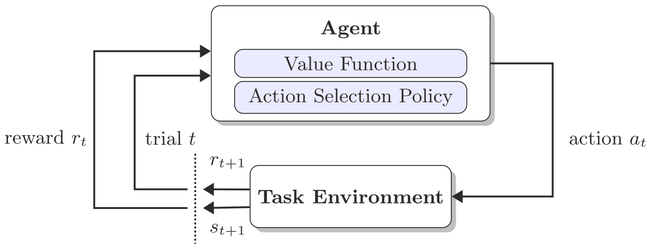


Fig. 2. The traditional framework of Multi-Armed Bandits.

shown in Eq. (1). Calling the arm with the highest expected reward at time t as the best arm, denoted as a_t^*, and its expected reward as the optimal reward r_t^*, the regret can be defined as the difference between r_t^* and the reward r_t achieved by the agent. Sometimes, it is also possible to redefine the objective function according to the work goal by maximizing the average reward returned, maximizing the percentage of optimal action selection, minimizing the number of trials without rewards, and so many others.

$$\text{Maximize} \underbrace{\sum_{t=1}^{T} r_t}_{\text{Reward}} \equiv \text{Minimize} \underbrace{\sum_{t=1}^{T} (r_t^* - r_t)}_{\text{Regret}} \qquad (1)$$

In general, the quality of any MAB algorithm is usually related to the way it handles the exploration and exploitation dilemma in its action selection policy π. If the agent only performs the exploration, it will choose the actions randomly by ignoring all the knowledge achieved in the earlier steps. On the other hand, if the agent only performs exploitation, it will choose the actions according to the short-term reward similarly to greedy approaches and, perhaps, never finding the long-term reward. After all, each agent's decision has long-term consequences: each action influences the environment and determines what type of information the agent can observe to update its policy going forward (Barraza-Urbina, Koutrika, d'Aquin, & Hayes, 2018). In this sense, many MAB algorithms have been proposed in the literature with different properties (Sutton & Barto, 2018). The main algorithms are called as ε-Greedy (Auer, Cesa-Bianchi, & Fischer, 2002), Upper Confidence Bounds (UCB) (Auer, 2002; Auer et al., 2002), and Thompson Sampling (TS) (Chapelle & Li, 2011). As these algorithms can exploit the same value function $Q_t(a) = \mathbb{E}[r_t|a]$, the main difference between them is related to the exploration step. In terms of the exploration strategies, these algorithms can: (1) do no exploration at all, focusing on the short-term returns; (2) occasionally explore at random; or (3) choose about which options to explore by favoring actions with higher uncertainty because they can provide higher information gain. These types of strategies usually classify the MAB algorithms. The next subsections show more details about each algorithm.

2.1. ε-Greedy

The ε-Greedy algorithm handles the exp–exp dilemma by selecting the best action most of the time and doing a random exploration occasionally. The best action is usually estimated according to the past experience by averaging the rewards associated with the target action a that was observed so far, as defined by Eq. (2). The value function Q for each arm a in the trial t usually considers the mean of rewards achieved in the earlier trials $\tau < t$. $N_t(a)$ is the number of times that the action a was taken before the trial t. This exploitation step is performed with probability $(1 - \varepsilon)$.

$$a_t^* = argmax_{a \in \mathcal{A}}\ Q_t(a) = argmax_{a \in \mathcal{A}} \frac{1}{N_t(a)} \sum_{\tau=1}^{t-1} r_\tau \qquad (2)$$

Otherwise, the algorithm performs the random exploration uniformly with probability ε. The main idea behind this step is to guarantee the algorithm to not get stuck in a suboptimal reward forever. However, the challenge is to define the ε. In general, this parameter gives a poor performance at the extremes. If it is too small, the learning ability defined by the exploration is slow at the start, and the algorithm will be slow to react to changes. In turn, if it is too big, the algorithm will waste many trials pulling random arms without gaining much. The best parameter should allow the algorithm to choose the best action for a large proportion of the time. Unfortunately, due to the randomness, the algorithm may end up exploring a bad action which the agent has already confirmed in the past. To avoid such inefficient exploration, there are two main approaches available. The first one is to decrease the parameter ε in time. The second and most usual one is to be optimistic about options with high *uncertainty* and, thus, to prefer actions for which the agent has not had a confident value estimation yet. This kind of exploration is considered smarter than the other because it favors the exploration of actions with a strong potential to have an optimal value. It is the main base of the other two strategies usually applied by MAB algorithms.

2.2. Upper Confidence Bound (UCB)

The Upper Confidence Bounds algorithm measures the function value $Q_t(a)$ by considering a confidence bound of the reward value, called as $C_t(a)$, so that the true value is below with bound $Q_t(a) \le Q_t(a) + C_t(a)$ with high probability. The main idea is to apply the upper bound to measure the potential of each action (arm) according to the uncertainty about its quality. Then, the agent always selects the greediest action with the highest upper confidence bound:

$$a_t^* = argmax_{a \in \mathcal{A}}\ Q_t(a) + C_t(a) \qquad (3)$$

In the literature, there are several approaches to measure the confidence bound $C_t(a)$ (Li, Chu, Langford, & Schapire, 2010; Li, Karatzoglou and Gentile, 2016; Nguyen & Lauw, 2014; Yu, Fang, & Tao, 2016). In general, $C_t(a)$ is defined as a function of $N_t(a)$, where a larger number of trials $N_t(a)$ is directly related to a smaller bound $C_t(a)$. The traditional UCB algorithm measures $C_t(a)$ by the Hoeffding's Inequality (Duchi, 2017), a theorem to any bounded distribution. Applying it, these works found that the probability of the expected reward being greater than the confidence bound is very small: $\mathbb{P}[Q(a_t) > \widehat{Q_t}(a) + C_t(a)] \le e^{-2tC_t(a)^2}$. In this sense, the main works define a very tiny threshold p and apply it to measure the confidence bound, as follow: $C_t(a) = \sqrt{-\log p\ /\ 2N_t(a)}$. One current heuristic aims to reduce the threshold p in time t to propose a parameter free algorithm. The famous $UCB1$ algorithm (Auer et al., 2002) is proposed setting $p = (t - 4)$ to make more confident bound estimation with more rewards observed. Its algorithm performs as follow:

$$C_t(a) = \sqrt{\frac{2 \log t}{N_t(a)}} \quad \text{and} \quad a_t^* = argmax_{a \in \mathcal{A}}\ Q_t(a) + \sqrt{\frac{2 \log t}{N_t(a)}} \qquad (4)$$

Furthermore, there are other works designed to explore the prior distribution of rewards by modeling the expected mean reward as a Gaussian and setting the upper confidence bound C according to the standard deviation. These approaches are called Bayesian UCB and they would be able to make better-bound estimation (Mahajan, Rastogi, Tiwari, & Mitra, 2012; Wang, Wang, Hsu, & Wang, 2014). A traditional implementation draws the reward distribution according to a prior distribution Beta with parameter θ and sets $C_t(a) = \alpha \cdot \sigma_t(a)$, where α is a adjustable hyperparameter and σ is the standard deviation of times a was chosen.

2.3. Thompson Sampling (TS)

Distinct from the other approaches, the Thompson Sampling algorithm implements the idea of probability matching. At each time step, the main idea is to select action a according to the probability that a

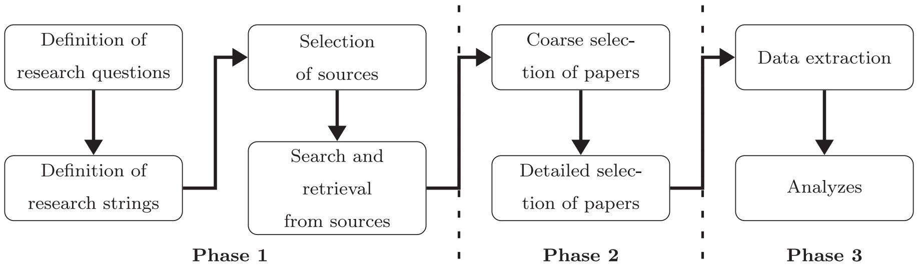

Definition of research questions
Definition of research strings
Selection of sources
Search and retrieval from sources
Coarse selection of papers
Detailed selection of papers
Data extraction
Analyzes
Phase 1
Phase 2
Phase 3

Fig. 3. Systematic literature review protocol.

is optimal according to the history of actions h_t already known by the agent:

$$\pi(a|h_t) \ = \ \mathbb{P}[Q(a) > Q(a'), \forall a' \neq a|h_t] \ = \ \mathbb{E}[a = argmax_{a\in\mathcal{A}}\ Q(a)] \quad (5)$$

Nowadays, the main TS algorithms are based on the Bernoulli bandit where they naturally assume that $Q(a)$ follows a Beta distribution by defining $Q(a)$ as the success probability θ in Bernoulli distribution. In general, the value of $\mathsf{Beta}(\alpha,\beta)$ is within the interval $[0,1]$ and the parameters α and β correspond, respectively, to the counts when the agent achieves success or failure to get a reward. Then, at each time t, the agent must sample an expected reward, $\tilde{Q}(a)$, from the prior distribution $\mathsf{Beta}(\alpha_i,\beta_i)$ for every action. The best action is selected among the samples: $a_t^* = argmax_{a\in\mathcal{A}}\ \tilde{Q}(a)$. After the true reward is observed, the Beta distribution is updated accordingly by doing Bayesian inference. If the action selected is correct (i.e., $a_t^* = a_i$), then α is updated. Otherwise, the β is incremented. This simple idea has worked very well in many scenarios (Chapelle & Li, 2011). However, for many practical and complex problems, it can be computationally intractable to estimate the posterior distributions with observed true rewards using Bayesian inference. In this case, it is necessary to approximate the posterior distributions using methods like Gibbs sampling, Laplace approximate, and bootstraps.

3. The systematic literature review protocol

A systematic literature review (SLR) is a scientific methodology designed to answer some well-formulated research questions. It aims to identify and synthesize all of the scholarly research on a particular topic by applying a rigorous, unbiased, and reproducible protocol. In general, there is a standard protocol usually defined by several steps at a high level to not consider the influence of research question type on the review procedures. Here, we design a protocol inspired by Çano and Morisio (2017) to assist the management of the review process in three main phases. First, we perform the paper collection by defining the search strings and searching in the main sources. Then, in phase 2, we perform the selection of papers in two steps: (1) a coarse selection, analyzing the title, year, conference, and the abstract; and (2) a detailed selection of the publications, reading the full paper. Finally, in phase 3, we extract the main information of the papers according to our criteria. Fig. 3 presents an overview of each step performed by this SLR protocol, representing a clear set of steps that are further discussed in the next sections.

3.1. Phase 1: Research questions, search strings and digital sources

First of all, this SLR defines the research questions to be applied in the main digital sources to identify and collect the most relevant papers in the literature. These questions are responsible to guide all the processes of our systematic literature review. Here, as this work is concerned about the application of Multi-Armed Bandits in the recommendation field, we design three main research questions to be answered by this SLR:

Q1: How have MAB algorithms been used in the RS field?

Q2: How have MAB algorithms been empirically evaluated in the RS field?

Q3: How have MAB algorithms dealt with classic challenges of the RSs literature, such as sparsity, privacy, cold-start, and explainability?

Each question is related to one of our three objectives. The first question is more inclusive and it may refer to most papers in the recommendation field related to MAB. Answering Q1, it is possible to consolidate a summary of the main research conducted in this area in the last years. In turn, the second question is more restrictive and it allows us to achieve the second goal. Q2 refers to papers of MAB in the recommendation field that have performed an experimental evaluation. Analyzing these papers we can identify how some MAB concepts are applied, what are the main adopted methods, what are their characteristics, and even how they are evaluated in this scenario. These analyzes can also drive several future directions for the literature. Finally, the third question allows us to achieve our third goal by providing insights about how MAB has been applied to face some traditional RSs' challenges. In Q3, we inspect papers related to bandit algorithms concerned with challenges such as the sparsity, privacy, cold-start, and explainability.

In order to answer these questions, we propose to search for publications written in English and published in the last two decades, from 2000 until 2020, into the main well-known repositories, such as IEEE Explorer, ACM Digital Library, Springer, Scopus, Scielo, and others. We advocate the use of the Google Scholar search engine because it automatically searches a query in these repositories. Therefore, we followed the protocol described on Keele et al. (2007) and applied the PICOC (Population, Intervention, Comparison, Outcome, Context) criteria to generate our search string (SS). In this work, the *population* is the Recommendation System literature, the *intervention* is the applicability of MAB algorithms, and the *outcomes* are used, evaluation criteria and the way how classic RS challenges are addressed. This string will be searched in the titles, highlights, and abstracts of each paper published so far.

```
SS: ("multi-armed bandit" OR "multi-armed bandits") AND
("recommender system" OR "recommender systems" OR "recommendation")
```

3.2. Phase 2: Paper selection

Applying this search string in the main sources, we identified 1327 articles by eliminating duplicates — those with the same title and publication link. To objectively decide whether to select each preliminary study for further processing or not, we defined a set of inclusion and exclusion criteria listed in Table 1. In general, we decided to include only journal and conference papers, leaving out gray literature, workshop presentations, or papers that report abstracts or presentation slides. Then, the selection of the most relevant studies was performed as follows. We initially analyzed title, publication year, and publication type (i.e., journal, conference, workshop, etc.), dropping out every paper that does not perform a recommendation study or does not address a bandit model at all. After this step, we reached a list of 408 papers (30.75%). These papers were then examined in a more detailed analysis by reading their introduction, experimental setup,

Table 1
Inclusion and exclusion criteria.

Inclusion criteria
- Papers presenting MAB algorithms in the context of recommendation systems. - Papers that even though do not specifically propose a new MAB algorithm, but contrast those MAB with other methods to perform online recommendations. - Papers from conferences and journals - Papers published from 2000 to 2020 - Papers written in English
Exclusion criteria
- Papers not addressing recommender systems at all - Papers addressing RSs but not considering a multi-armed bandit modeling - Papers that report only abstracts or slides of presentation, lacking detailed information - Gray literature

and conclusion. This process was made by three people who read at least 2/3 of the papers selected and shared their knowledge to decide the most important ones. Thus, each paper was read for at least two distinct people who debated its characteristics and got a consensus about their relevance for this study. From the 408 papers, 178 papers were rejected due to the same criteria listed, and 230 papers were selected as relevant studies for our SLR. These studies represent the most relevant works about MAB in the recommendation field and they are the focus of the remaining discussion of this work.

3.3. Phase 3: Data extraction

In order to achieve our three main goals, we also propose to collect the main information for each paper. Then, we built a data extraction form to collect both paper metadata (i.e., author, title, year, etc.) and the relevant content data for answering our research questions. This relevant content is defined based on the needed information to develop, apply, and analyze any MAB method in the field. It is also built to provide an overview of the main published works. The complete form and intuition of each data extracted are described in Table 2. The extraction process happens during the second step of reading when we examined the introduction, experimental setup, and conclusion of each paper. Here, we focused on 190 papers from the 230 works previously selected due to their experimental nature (i.e., those that performed experimental analyzes). Moreover, we discussed each paper with each other to answer our doubts.

4. Multi-Armed BAndits in the recommendation field

Nowadays, several works have modeled the online recommendation task as a Multi-Armed Bandit problem (Felício, Paixão, Barcelos, & Preux, 2017; Wang, Wang, Wang and He, 2017; Wang et al., 2018). In most of the bandit representations, the items to be recommended are modeled as the arms to be pulled. Selecting an arm a is equivalent to recommending an item i and the reward is the user response to this recommendation (e.g., clicks, ratings, acceptance, etc.) (Sanz-Cruzado et al., 2019). Thus, the main goal is also to maximize the expected reward achieved after T times, as shown in Eq. (6). Here, however, the difference from traditional learning scenarios is the goal, which must be related to the user's satisfaction with the system. It requires a personalized action selection policy π to the users' preferences and tastes identified by the historic of user's actions h. For this reason, the item i_t^* should be chosen according to a prediction rule π, which is defined as a function to exploit and explore the current known information about the user until now: $i_t^* \equiv \pi(h_t)$. In the literature, there are a lot of distinct strategies to define and improve this action selection policy π.

$$i^*_{(\cdot)} = \underset{i_{(\cdot)}}{argmax} \sum_{t=1}^{T} \mathbb{E}\left[r_{u,i_t}|t\right] \tag{6}$$

In this sense, the primary goal of our systematic literature review is to shed light on the current advances of MAB in the recommendation field. We intend to provide an overview of the most used concepts and methods, their characteristics, and also how they can be applied, evaluated, or improved. We guide our analysis in the 190 papers previously selected to answer the three research questions raised before. First, we highlight what the current works have proposed in the last few years by exploring some data extracted, such as the domain, dataset, method, RS characteristics, and context. Next, we analyze the evaluation criteria of MAB in the recommendation field by inspecting the experimental setup of the selected publications. Here, we investigate specific characteristics of the papers, such as the data type, the data processing approach, the evaluation metrics applied, and also the methodology chosen by the authors. And, finally, we discuss the application of MAB in traditional RSs' challenges by highlighting the number of works concerned with usual problems and their assumptions to handle each of them. Furthermore, the analyzes of these three questions also allow us to identify several directions for future research about MAB in recommender systems.

4.1. The works developed so far

The application of MAB in the recommendation field has received more attention recently. Despite the recommendation field had emerged in the mid of 90 s, the first studies about this topic started in 2005 and it has become more relevant after 2010. Especially, Fig. 1 shows that more than 50% of all works published about MAB in RSs were in the last five years (2016 to 2020). The main explanation for this current interest of the literature may be related to two main factors: the saturation of methods and concepts in the traditional scenario due to the great advances made in the last decade; and the increasing investment of several companies in RSs by requiring models capable to deal with the users' needs without having to retrain the prediction model at each interaction. In general, both factors have influenced several conferences to call for papers about the MAB application

Table 2
Data extraction form.

Data	Explanation	Examples
Title	–	–
Authors	–	–
Publication year	–	–
Conference	–	–
Source	Digital library source	–
Domain	Domains in which the study is applied	News, movies, songs, ads.
Dataset	Public or private dataset used to train the algorithm	Yahoo!, MovieLens, Netflix.
Method	The bandit method applied in the study	ϵ-Greedy, UCB, TS.
RS Characteristics	The recommendation concept applied to guide the bandit algorithm	Graph-based, matrix factorization, probabilistic.
Context	A context applied to improve the recommendation	User profile, items' characteristics, cookies.
Data Type	The type of feedback provided by the user	Rating, clicks, likes.
Data Processing	A processing stage performed to normalize the data	Mean-centering, Z-score.
Metrics	The main metrics applied to measure the recommendation's quality	Precision, recall, regret, CTR.
Evaluation	The methodology applied to simulate the online user's interaction	Trials, interactions.

Table 3
Conferences and journals that have accepted more papers about multi-armed bandits in the recommendation field.

#	Conference/Journal	Papers	#	Conference/Journal	Papers
1	RecSys	21	11	IJCAI	4
2	WWW	12	12	KDD	4
3	NIPS	11	13	TKDE	3
4	ICML	10	14	SAC	3
5	CIKM	9	15	University Lib	3
6	SIGIR	8	16	ICTAI	3
7	AAAI	7	17	UMAP	3
8	AISTATS	6	18	UAI	3
9	ICONIP	4	19	JMLR	3
10	WSDM	4	20	Neurocomputing	3

in the recommendation field and propose interesting talks between researchers and business leaders of big companies such as Netflix, Spotify, Google, Pandora, and others. Here, in our SLR, we identified 112 distinct conferences/journals that have accepted at least one paper about this topic. Table 3 shows an overview of these conferences and journals by highlighting the top-20 that have accepted more papers. Therefore, the first insight we found is that MAB has been extensively discussed in the main conferences around the world.

Moreover, we identified that several approaches have been proposed for distinct domains by simulating a recommendation of movies to watch, news to read, products to buy, songs to hear, and others. Specifically, in the 190 papers with experimental evaluations previously selected by our SLR, we identified 50 distinct domains highlighted according to the number of times each one is applied in a paper (i.e., their frequency) in Fig. 4. In general, the recommendation of Movies and News are the most usual (Celis, Kapoor, Salehi, & Vishnoi, 2019; Liu, Wei, Zhang, Yan and Yang, 2018; Rao, 2020; Silva, Silva, Werneck, Pereira, & Rocha, 2020; Song, Fragouli, & Shah, 2019; Tracà, Rudin, & Yan, 2020; Wu, Wang, Hong, & Shi, 2017; Zhang, Xie, Li and Lui, 2020; Zhou et al., 2020) by appearing in 24.7% and 22.1% of all papers respectively. The main works usually simulate these domains with traditional datasets available in the literature, such as the MovieLens library, the collection of news from the Yahoo front page, and the Amazon repository of users' purchases. Moreover, it is also a usual practice to evaluate MAB algorithms in synthetic or generic domains. While the synthetic domain is usually designed to simulate random (i.e., without bias) interactions of users with items, the generic domain refers to traditional datasets often applied to evaluate some Machine Learning algorithms (e.g., Iris Flower Species,[2] CNAE-9,[3] and others). In addition, we also noticed: (1) works proposed to other traditional recommendation domains, like Songs (Jagerman, Markov, & de Rijke, 2019; McInerney et al., 2018; Takemori, Sato, Sonoda, Singh, & Ohkuma, 2020; Tripathi, Ashwin, & Guddeti, 2018; Wang, Wu et al., 2017; Wu et al., 2020, 2016; Zhou, Jin, Wang and Zhang, 2020), Artists (Caron & Bhagat, 2013; Gentile, Li, & Zappella, 2014; Hariri, Mobasher, & Burke, 2014; Nguyen & Lauw, 2014; Xu, Dong, Li, He, & Li, 2020; Zhao & King, 2016), Ads (Aharon, Kagian, Kaplan, Nissim, & Somekh, 2015; Gentile et al., 2017; Li, Karatzoglou et al., 2016; Martín, Jiménez-Martín, & Mateos, 2019; Tang, Jiang, Li, & Li, 2014; Tang, Jiang, Li, Zeng, & Li, 2015; Warlop, Lazaric, & Mary, 2018), Points-of-Interest (POIs) (Chen, Xu, & Lu, 2018; Gutowski, Amghar, Camp, & Hammoudi, 2017; Gutowski, Camp, Chhel, Amghar, & Albers, 2019; Sanz-Cruzado et al., 2019; Song, Fragouli, & Shah, 2018; Theocharous, Vlassis, & Wen, 2017; Wang, Liu, Jiang, Li and Fu, 2020; Zong et al., 2016), Products (Brodén, Hammar, Nilsson, & Paraschakis, 2018; Goswami, Zhai, & Mohapatra, 2019; Peng et al., 2019; Wen, Wang, Wu, Liu, & Cao, 2020; Yan et al., 2018; Yu, Shen, & Jin, 2020), and Bookmarking (Cesa-Bianchi, Gentile, & Zappella, 2013; Gentile et al., 2017, 2014; Wang et al., 2020; Wu et al., 2020); (2) works that apply recommendation MAB algorithms to offer personalized rankings in the traditional Information Retrieval task (Vorobev, Lefortier, Gusev, & Serdyukov, 2015; Yue & Guestrin, 2011); and (3) works especially focused on non-usual scenarios like Jokes (Kong, Brunskill, & Valiant, 2020; Rahman & Oh, 2015a, 2015b, 2018; Xu, Vakili, Zhao, & Swami, 2017), Food (Gutowski, Amghar, Camp, & Chhel, 2019b; Gutowski, Camp, Amghar, & Chhel, 2019c; Immorlica, Mao, Slivkins, & Wu, 2019), and Jobs (Gomes, Almeida, & Vale, 2020; Sato, Nagatani, & Tahara, 2017).

However, despite the huge number of distinct domains currently used, the algorithms applied to handle the MAB problem are often related to the same approaches. Extracting the data related to ′Method′ of each paper previously selected by our SLR, we calculate in our analysis the percentages of works related to each MAB approach in Fig. 5(a). As we can see, the most used MAB algorithms are based on UCB, Thompson Sampling and ε-Greedy approaches, with 45.8%, 18.9%, and 16.8% of works, respectively. Other approaches, such as P2EE (Ravi, Poduval, & Moharir, 2020) and EXP3 (Li, 2011; Louëdec, Chevalier, Mothe, Garivier, & Gerchinovitz, 2015; Nguyen & Kofod-Petersen, 2014) are unusual and less effectively. The works that are not concerned with a specific MAB algorithm, giving the option to apply any learning algorithm, are classified as independent from MAB ('i.i.d.MAB'). These works were not excluded by our SLR since we are also interested in all spectrum related to applications of MAB. Our SLR also found works focused on designing other learning algorithms for specific scenarios that are classified here as 'Others'. Besides, Fig. 5(b) highlights the application evolution of the three main MAB approaches over the last five years. Despite the greater number of publications based on the UCB approach, we can notice a growing interest among researchers about Thompson Sampling (TS) recently. The main explanation can be associated with the lack of theoretical analysis about TS before the publication of the paper named 'An Empirical Evaluation of Thompson Sampling' (Chapelle & Li, 2011). After this relevant work, more researchers could propose new MAB algorithms based on TS approaches, reflecting it in the number of publications in the next years.

Similarly, extracting the data related to 'RSCharacteristic' of each paper previously selected by our SLR, our analyzes also find the percentages of works where the RSs techniques are applied and also the number of techniques by year. A picture of this information is shown in Fig. 6. Here, these techniques are strongly relevant for the MAB algorithms because they must be combined to improve the

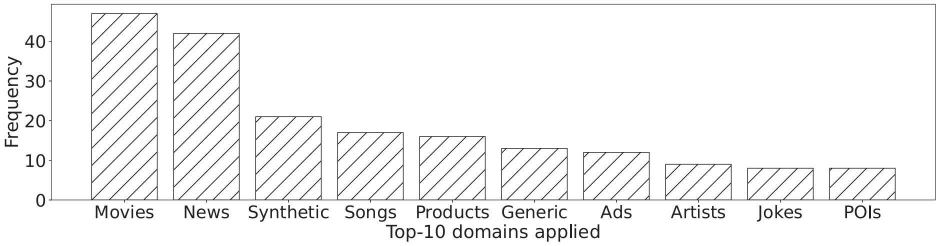
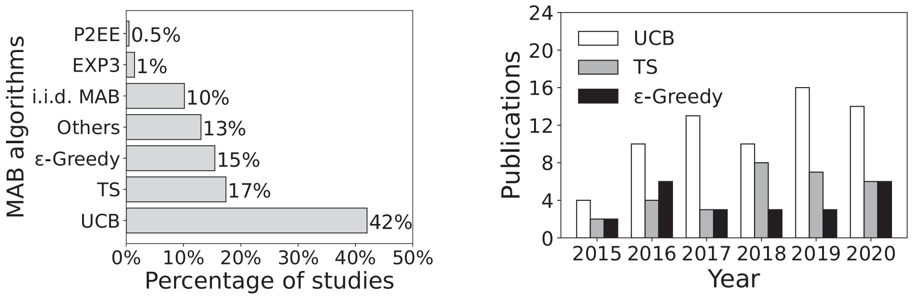


Fig. 4. The main domains used in the recommendation field to apply MAB.

[2] http://archive.ics.uci.edu/ml/datasets/Iris.

[3] https://archive.ics.uci.edu/ml/datasets/CNAE-9.



(a) MAB approaches

(b) MAB approaches in last five years

Fig. 5. A picture of the MAB approaches usually applied in the recommendation field.

recommendation's quality (Zhao, Xia et al., 2019). The most usual RSs techniques aim to identify relevant items with distinct assumptions. Some techniques have assumed that users will like items that were liked by other users with the same preferences and tastes, designing algorithms based on Clusters (Bagaria, Kamath, Ntranos, Zhang, & Tse, 2018; Bouneffouf & Claeys, 2016; Chen, Krause, & Karbasi, 2017; Chi, Lin, & Ing, 2019; Gentile et al., 2017, 2014; Geyik, Dialani, Meng, & Smith, 2018; Jedor, Perchet, & Louedec, 2019; Khenissi, Mariem, & Nasraoui, 2020; Li, Karatzoglou et al., 2016; Song, Tekin, & Van Der Schaar, 2014a; Yang, Li, Qin and Ye, 2020; Zhou, Wang, Guo, Gong, & Zheng, 2019), Graphs (Bouneffouf, 2016; Bouneffouf, Bouzeghoub, & Gançarski, 2012a; Cesa-Bianchi et al., 2013; Li, Jiang, & Li, 2017; Li, Karatzoglou et al., 2016; Wu et al., 2016; Zhao, Watanabe, Yang, & Hirate, 2018), and Matrix Factorization (Brodén, Hammar, Nilsson, & Paraschakis, 2019; Guillou, Gaudel, & Preux, 2016; Gupta, Balaji, & Luo, 2020; Li et al., 2010; Maniu, Ioannidis, & Cautis, 2020; Mehrotra, Xue, & Lalmas, 2020; Mukherjee, Kveton, & Rao, 2019; Silva et al., 2020; Tavakol & Brefeld, 2017; Tekin & Van Der Schaar, 2015; Teo et al., 2016; Wang, Wu et al., 2017; Zhao et al., 2013). Other ones have considered some statistical fundamentals to model the user's behavior and, then, developed algorithms based on Bayesian (Christakopoulou, Radlinski, & Hofmann, 2016; edw, 2020; Geyik et al., 2018; Hsieh, Neufeld, King, & Cho, 2015; Mahajan et al., 2012; Wang et al., 2014; Zhao, Yang and Hirate, 2019) or Probabilistic concepts (Balakrishnan, Bouneffouf, Mattei, & Rossi, 2019; Bernardi, Estevez, Eidis, & Osama, 2020; Bouneffouf, Rish, Cecchi, & Féraud, 2017; Dumitrascu, Feng, & Engelhardt, 2018; Hariri et al., 2014; Liu, Cai, & Zhang, 2017; Manickam, Lan, & Baraniuk, 2017; Mishra & Thakurta, 2014; Saritaç & Tekin, 2017; Tang et al., 2014; Wang et al., 2017; Wu et al., 2019; Yu, Mengshoel, Meroux and Jiang, 2019; Yu, Shen and Jin, 2019; Zeng, Wang, Mokhtari, & Li, 2016; Zhang et al., 2020; Zhang, Zhou, He, & Liang, 2018; Zhu, Huang, & Xu, 2020b). Our SLR also identifies other technique, related to memory-based (Bouneffouf, 2016; Bouneffouf et al., 2012a; Bouneffouf, Bouzeghoub, & Gançarski, 2012c; Bresler, Chen, & Shah, 2014; Heckel & Ramchandran, 2017; Matikainen, Furlong, Sukthankar, & Hebert, 2013; Sanz-Cruzado et al., 2019; Tekin & Van Der Schaar, 2015; Zhong, Ying, Chen, & Fu, 2020) and hybrid approaches (Felício et al., 2017; Gutowski et al., 2019c; Lacerda, 2017; Lacerda, Veloso, & Ziviani, 2015; Liang, Loni, & Larson, 2017; Lu, Wen, & Kveton, 2018; Santana et al., 2020; Tang et al., 2014; Wu, Wang, Li and Wang, 2019). Moreover, it is important to observe that the three most applied techniques are based on Collaborative Filtering (CF) concepts, the most popular class of RSs in the literature (Bobadilla et al., 2013). Therefore, we also analyze the total of works that consider these techniques over the years in Figs. 6(b). This figure shows the current growth of interest in Probabilistic and Matrix Factorization (MF). Indeed, these techniques have achieved the best results in the last years.

Moreover, Fig. 7 highlights an updated picture of how MAB algorithms have been developed in the literature. It correlates the MAB approaches with the top-10 RSs techniques used to indicate how the action selection policy π has been implemented by the most relevant works in the field. The numbers show how many works were published with a MAB approach combined with an RSs strategy. In addition to the observations made before, we can also notice the influence of Collaborative Filtering concepts in MAB algorithms. From the top-10 RSs techniques applied, six of them are memory-based or model-based approaches. And, for the best of our knowledge, even the hybrid and heuristic approaches usually apply a collaborative filtering representation combined with other information.

Then, studying this majority set of MAB algorithms (those CF-based), we identified two main groups of works with the same assumption of similarity between arms (i.e., playing one arm will give you information about similar arms). The first one has described the reward structure in terms of clusters of users and/or items and defined it similarly to memory-based methods (e.g., the k-Nearest Neighbors) (Gentile et al., 2017, 2014; Li, Karatzoglou et al., 2016). In turn, the other one has assumed the similarity structure between arms by constructing a model for each arm similar to CF model-based methods (Wang, Wu et al., 2017; Yue & Guestrin, 2011; Zhao et al., 2013). While some works have assumed a non-linear model applying a deep learning algorithm, other works have assumed a linear model and represented items/users probabilistically by vectors of features extracted from MF approaches (Li et al., 2010; Wang, Wu et al., 2017; Wang, Zeng

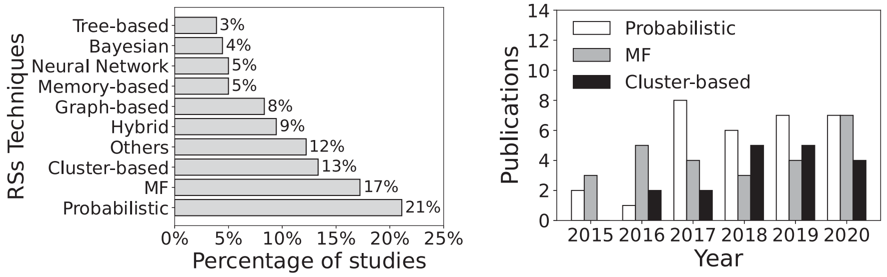


Fig. 6. A picture of the RSs techniques usually combined with MAB algorithms.

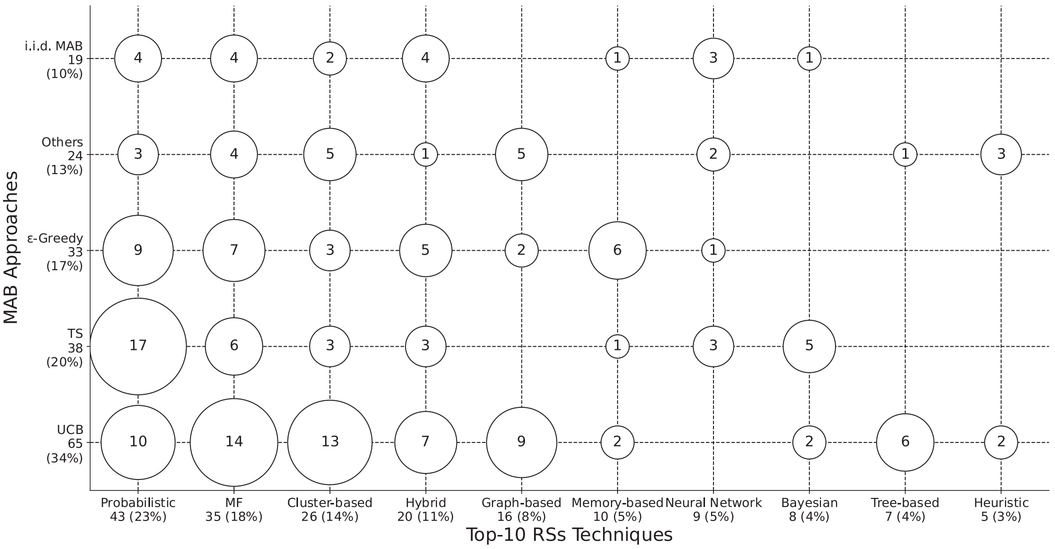


Fig. 7. An updated picture of the main RSs' techniques applied with MAB approaches according to the works selected by our SLR.

et al., 2018; Zhao et al., 2013). The linear bandits are most frequently applied and they are usually represented in two ways. Some works represent the reward by a Bernoulli (Cao, Wen, Kveton, & Xie, 2019) or Gaussian (Hariri et al., 2014; Hariri, Mobasher, & Burke, 2015; Wang, Li, Iyengar, Shwartz and Grabarnik, 2018) distribution, applying (or not) a Bayesian Inference. They model an unknown probability distribution over the rewards ($\mathcal{R}^i(r) = \mathbb{P}[r|i]$) and try to identify the arm that will maximize the expected reward returned. In turn, other linear works usually represent each item i and user u by their features vector extracted from an MF application called as $\boldsymbol{q_i}$ and $\boldsymbol{p_u}$ respectively. Here, the rating matrix $M^{m\times n}$ is estimated according to the product of two low-rank matrices $P \in \mathbb{R}^{m\times z}$ and $Q \in \mathbb{R}^{z\times n}$. While the matrix $P^{m\times z}$ contains the user-model p_u, representing the multiple interests of each user u in the z features, the matrix $Q^{z\times n}$ represents how relevant is the item i for the z groups. In this case, the expected reward is reformulated for the product of user and item features vectors:

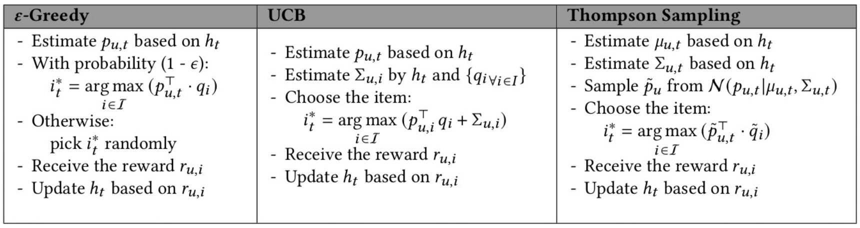
$$i^*_{(\cdot)} = \underset{i_{(\cdot)}}{argmax} \sum_{t=1}^{T} \mathbb{E}\,[r_{u,i_t}|t] = \underset{i_{(\cdot)}}{argmax} \sum_{t=1}^{T} \mathbb{E}\,[\boldsymbol{p}_{\boldsymbol{u}}^{\top} \boldsymbol{q}_{i_t}|t] \qquad (7)$$

In general, both CF approaches are extremely relevant but the model-based ones achieve better recommendations since they model non-trivial relations between users and items (Bobadilla et al., 2013). The difference in the interactive recommendation scenario is the system should learn with each user interaction. Then, the current efforts have been focused on how to optimize this objective function while balancing the exp–exp dilemma. Traditional MAB approaches, such as the ϵ-Greedy, UCB, and Thompson Sampling, have been adapted for this objective function (Chapelle & Li, 2011; Hariri et al., 2014; Li et al., 2010; Wang, Wu et al., 2017; Zhao et al., 2013). After conducting a detailed reading in the works selected by our SLR, we notice the main linear bandits have followed the same prediction rules of traditional MF algorithms. Thus, in Fig. 8, we could define a standard implementation for these three main MAB approaches by following a linear assumption. Again, the difference between ϵ-Greedy, UCB, and Thompson Sampling algorithms is more related to the way they control the exp–exp dilemma. UCB and Thompson Sampling algorithms perform a smart exploration by measuring an uncertainty Σ around the information available about the user and items. However, all approaches perform exploitation based on p_u and q_i previously defined according to the system's historic h_t available.

Furthermore, current approaches usually introduce the user/item context available in the system to provide more effective recommendations (Tekin & Turğay, 2018; Wu et al., 2019; Zhong et al., 2020). Such approaches are named contextual Multi-Armed Bandits (CMAB) and they have been studied in around 43% of works selected by our SLR. Basically, in a typical contextual bandit setting, each arm a is associated with a d-dimensional context vector x_a, and its expected reward is guided by a conjecture of the context vector and an unknown bandit model, parameterized as θ^*. Here, in the personalized online recommendation field, the unknown bandit model parameter θ^* is usually attached to each user to reflect their corresponding personalized preferences where θ^*_u for $u \in \mathcal{U}$ are independently estimated based on the observations from the corresponding users. Then, in a linear contextual bandit setting, it is assumed that $r_a \sim \mathcal{N}(x_a^\top\theta^*, \sigma^2)$ and the expected reward r_t achieved in a specific context x_a is defined as $\mathbb{E}[r_t|x_a] = x_a^\top\theta^*$. One of the most famous CMAB algorithms was designed by following these assumptions to maximize the total reward (i.e., the total user clicks in the long run). It is named as LinUCB (Li et al., 2010) and it measures upper confidence bound for each item according to the context x for choosing one of them at every trial t:

$$i^*_t = \underset{i\in\mathcal{I}}{argmax}\,(x_{t,i}^\top\,\theta^* + \alpha\sqrt{x_{t,i}^\top \mathbf{A}_i^{-1} x_{t,i}}) \qquad (8)$$

In general, the context x often summarizes information of both the user u_t and the item i (kam, 2011; Liang et al., 2017). However, our



ϵ-Greedy	UCB	Thompson Sampling
- Estimate $p_{u,t}$ based on h_t - With probability (1 - ϵ): $i^*_t = \arg\max_{i\in\mathcal{I}} (p_{u,t}^\top \cdot q_i)$ - Otherwise: pick i^*_t randomly - Receive the reward $r_{u,i}$ - Update h_t based on $r_{u,i}$	- Estimate $p_{u,t}$ based on h_t - Estimate $\Sigma_{u,i}$ by h_t and $\{q_{i\forall i\in\mathcal{I}}\}$ - Choose the item: $i^*_t = \arg\max_{i\in\mathcal{I}} (p_{u,i}^\top q_i + \Sigma_{u,i})$ - Receive the reward $r_{u,i}$ - Update h_t based on $r_{u,i}$	- Estimate $\mu_{u,t}$ based on h_t - Estimate $\Sigma_{u,t}$ based on h_t - Sample $\tilde{p}_u$ from $\mathcal{N}(p_{u,t}\|\mu_{u,t}, \Sigma_{u,t})$ - Choose the item: $i^*_t = \arg\max_{i\in\mathcal{I}} (\tilde{p}_{u,t}^\top \cdot \tilde{q}_i)$ - Receive the reward $r_{u,i}$ - Update h_t based on $r_{u,i}$

Fig. 8. The standard linear implementation of the three main MAB algorithms.

Table 4
The main repositories with the datasets most applied by researchers to study MAB proposals in the recommendation field.

Repository	Domain	Available at
MovieLens	Movies	https://grouplens.org/datasets/movielens/
Yahoo! News	News	https://webscope.sandbox.yahoo.com/
LastFM	Music	https://grouplens.org/datasets/hetrec-2011/
Delicious	Bookmarking	https://grouplens.org/datasets/hetrec-2011/
Netflix	Movies	https://kaggle.com/netflix-inc/netflix-prize-data
Jester	Jokes	https://goldberg.berkeley.edu/jester-data/
KDD - Online Ads	Advertising	https://www.kaggle.com/c/kddcup2012-track2/overview
Avazu	Advertising	https://kaggle.com/c/avazu-ctr-prediction
Amazon	Products	http://jmcauley.ucsd.edu/data/amazon/links.html
Yahoo! Music	Music	https://webscope.sandbox.yahoo.com/catalog.php?datatype=r
Millionsong	Music	http://millionsongdataset.com/
Yelp	Restaurant	https://yelp.com/dataset_challenge
RS-ASM	Smart Cities	https://kaggle.com/assopavic/recommendation-system-for-angers-smart-city
Epinions	Products	http://alchemy.cs.washington.edu/data/epinions/

SLR also finds several definitions of context in the selected works. Some works simply model the d-dimensional vector x as a traditional embedding of TF-IDF values, measuring it by the items and users' tags (Gentile et al., 2014; Wu, Iyer, & Wang, 2018). Currently, the works have modeled the context as users/items feature vectors extracted from an MF formulation (edw, 2020; Hariri et al., 2014, 2015; Wang, Hoi, Liu and Ester, 2017; Wang, Zeng et al., 2018; Zhao et al., 2013). Other works, in turn, define the context with other information available in the dataset. Song, Tekin, and Van Der Schaar (2014b), Bouneffouf, Bouzeghoub, and Gançarski (2012b), and Gutowski et al. (2019b), for example, use all information available about the users as context. They propose an embedding with the cookies of the users' last accesses to the system, their demographic information (e.g., gender, age, etc.), and any other information available in their profile (e.g., preferences, calendar, social connection, etc.). Other works, like (Bouneffouf, 2014; Li, Karatzoglou et al., 2016; McInerney et al., 2018), apply explicit information about the items to model the context. This type of data is very informative in specific domains, such as news (Li et al., 2010; Li, Karatzoglou et al., 2016; Tang et al., 2014), e-commerce (Wu, Zhang et al., 2019; Zhu et al., 2019), and POIs (Eide & Zhou, 2018; Zong et al., 2016). Moreover, there are other works focused on extracting the context of users/items according to their cluster of interests in the domain (Bouneffouf, Laroche, Urvoy, Féraud, & Allesiardo, 2014; Yang & Toni, 2018).

4.2. The evaluation criteria of Multi-Armed Bandits in the field

In the literature, the evaluation criteria is a protocol defined to measure if a recommendation technique is (or is not) effective in a domain (Chen & Liu, 2017; Cremonesi, Garzotto, Negro, Papadopoulos, & Turrin, 2011). Throughout the decades of recommendation systems research, it has been continually adapted and improved for several works (Castells, Vargas, & Wang, 2011; Vargas, 2011; Vargas & Castells, 2011) by performing two distinct criteria to measure the recommendation quality: (1) an offline experimentation; and (2) an online user's study. While offline experiments are often concerned about the prediction power – their ability to accurately predict the user's choices, the online assessments aim to measure the real business value – the actual value gain that a recommendation system can achieve to the system. However, due to perform AB tests with actual users, an online evaluation is more expensive than an offline and it usually requires a real system to collect the users' actions and opinions (Krohn-Grimberghe, Nanopoulos, & Schmidt-Thieme, 2010; Shani & Gunawardana, 2011). Hence, it is a common practice to identify more offline experimentations than online user studies, especially when it is studying new scenarios and applications.

Indeed, in our systematic review, studying the applicability of MAB in the recommendation field, we identified only 5 (2.6%) works that performed an online evaluation of their bandit algorithms. Once our study is really new, most of the works are still concerned into analyze their algorithms by offline experiments. By reviewing their evaluation criteria we identified three main steps to perform this offline evaluation, that will guide our discussion:

1. Selecting or creating a dataset to simulate the recommendation scenario;
2. Performing data processing when it is necessary; and
3. Measuring the main quality metrics based on the methodology used to simulate the item consumption.

These criteria are essential to provide a comparative evaluation of the main works and also provide guidelines for other MAB works.

4.2.1. Usual datasets

In the MAB application on the recommendation field, we identified distinct datasets that are selected or even created to measure the bandit performance. One usual practice is to create synthetic datasets to be applied specifically for some experimental analyzes. However, it is not encouraged because these datasets are unreproducible and cannot provide confidence in the results achieved. The applicability of real-world datasets is undoubtedly the best practice for analyzing the performance of any new recommendation technique. Fortunately, from the papers selected by our SLR, 186 works (80.9%) apply at least one real dataset in their experimentation. Our SLR highlights the datasets most applied to evaluate distinct bandit algorithms in Table 4 and also provides how other researchers can find them.

4.2.2. The data processing

In most of these datasets previously mentioned, there is not only the information available about the users and items of the system but also the feedback that was given by a specific user to an item of the catalog. These feedbacks are traditionally explicit, when the user says explicitly what s/he likes and dislikes, or implicit when it remains not clear if the user liked or not of some item. For this reason, similar to traditional recommendation scenarios, 19% of works selected by our SLR have performed a data processing step to clean the data available by removing incomplete data, noises, and redundant records. These works usually perform normalization of their rating data by mapping each rating according to static scales or specific functions. Some works, like (Cañamares, Redondo, & Castells, 2019), simply defined that ratings less or equal to 3 mean the user did not like the item (mapping as 0) and, otherwise, ratings greater than 3 mean the user likes this item (mapping as 1). Other works, such as the proposed in Basu, Sen, Sanghavi, and Shakkottai (2019), define a specific function to normalize their ratings by converting each value to a $[0,1]$ range. There is not a consensus about the better practice behind the normalization. However, it has become even more popular after the growth of UCB and Thompson Sampling applications due to their statistical assumptions of distributions such as the Beta and Bayesian (Zhao, Yang et al., 2019).

4.2.3. Methodologies & metrics

As the bandit algorithm is an online learning method (see Fig. 2), most of the offline experiments must define an approach to simulate the user interaction with a real system. In the papers selected by our SLR, we identified five main distinct approaches applied during the experimental evaluation:

- **Trials**: when a rating is estimated for one user–item pair in each interaction (Chatterji, Muthukumar, & Bartlett, 2020; Hao, Yadkori, Wen, & Cheng, 2019; Katariya, Kveton, Wen, & Potluru, 2019; Li, Wang, Zhang and Chen, 2016; May, Korda, Lee, & Leslie, 2012; Zhao, Yang et al., 2019)
- **Interactions**: when more than one item is recommended for the user (Christakopoulou & Banerjee, 2018; Liu, Li and Zhang, 2018; Sui, Gotovos, Burdick, & Krause, 2015; Teo et al., 2016; Vorobev et al., 2015; Wang, Wang et al., 2017)
- **Replayer:** when the system tries to predict the items consumed chronologically according to the user's consumption historic (Li, Chu, Langford, & Wang, 2011; Tang et al., 2014; Wang, Li et al., 2018; Wang, Zeng et al., 2018; Zeng et al., 2016)
- **Leave-one-out**: when the system left one item out of the training step and tries to predict this specific item for a given user (Felício et al., 2017)
- **Cross-validation**: when the system performs the same user interaction for many times by applying distinct kind of data at the training step (Hariri et al., 2014; Wang et al., 2014)

Based on our analyzes, the methodology most applied is the *trials*, by covering more than 45.3% of the works selected. The second one is the *Interactions*, being applied in 8.9% of the works. *Replayer*, *Leave-one-out*, and *Cross-validation* are only applied in 2.6%, 0.5%, and 1.1% respectively. Possible, the popularity of *trials* is related to a traditional bias that still exists in the evaluation criteria of several works. Similar to the first decade of researchers in the recommendation field, most bandit algorithms have been evaluated only according to their prediction power — their ability to accurately predict the user's choices. However, it is now widely agreed that accurate predictions are crucial but insufficient to deploy a good recommendation engine (Castells, Hurley, & Vargas, 2015; Castells et al., 2011). In many applications, people use a recommendation system for more than exact anticipation of their tastes. Users may also be interested in discovering new items, in rapidly exploring diverse items, in preserving their privacy, in the fast responses of the system, and many more properties of the interaction with the recommendation engine.

However, this new consensus is not a reality for MAB researchers yet. Indeed, most evaluation metrics applied are to measure traditional prediction power. Especially, the researchers brought the classical metrics of reward and regret the recommendation field and focused their bandit algorithms to maximize (minimize) it. Basically, these metrics represent the error in estimating a rating for a specific user–item pair. Other metrics like *precision*, *recall*, and *nDCG*, classical in the recommendation field, have been applied only for less than 20% of the papers. Fig. 9 shows the main metrics implemented by the works selected by our SLR.

4.3. Facing the recommendations challenges with MAB algorithms

Once the current works have combined MAB approaches with the main RSs techniques, their solutions are susceptible to face real challenges of the recommendation field. In this sense, our SLR also found many works concerned to address real RSs' challenges, such as data sparsity, cold-start, privacy, or explainability.

4.3.1. Data sparsity & scalability

Traditionally, while sparsity refers to the absence of data in a dataset, scalability refers to the difficulty of proposing an algorithm capable of presenting quick responses. In especial, in the recommendation field, both are strictly related because they are a consequence of the quick-growing of datasets according to more users and items are added in real-world systems. With more items in a system, the difficulty for users rating all of them also increases, worsening the data sparsity problem. Similarly, with more users in a system, it is more expensive to process all data and provide recommendations for all of them. Then, developing a scalable algorithm to handle the huge data sparsity is one of the great challenges of recommender systems.

In the literature, many works have been proposed to face this problem in traditional scenarios, but only a few of them have also been concerned about it in the bandit domain. In general, we identified scalability and sparsity being handled in distinct ways. Some works have proposed a cluster-based algorithm to address (or even summarize) the huge number of distinct dimensions (i.e., preferences) in a manageable number. The most relevant works in this sense are CLUB (Crammer & Gentile, 2011), DynUCB (Nguyen & Lauw, 2014), and COFIBA (Li, Karatzoglou et al., 2016). In particular, Nguyen and Lauw (2014) present studies of how to alleviate the sparsity problem with clusters of users. Other works, like (Wu, Zhang et al., 2019) propose to deal with sparse scenarios through social information available from items and users. Indeed, all of these techniques help to improve the scalability of MAB algorithms by training their models only for groups of users. On the other hand, there are other works focused exclusively on such a scalability problem. Rahman and Oh (2015b), for instance, proposes a parallel version to the classic UCB algorithm, proposed in Auer et al. (2002). And, Tekin, Zhang, and van der Schaar (2014) and Nguyen, Pham, Son, and Hong (2016) present distributed MAB algorithms for recommendation. Despite that, we did not find proposals that goals to deal with both challenges simultaneously.

4.3.2. Cold-start problem

The cold start problem is a well-known problem for recommender systems in the literature. It usually refers to the absence of information that makes it more difficult to provide personalized recommendations. In general, there are three cases of cold start (Bobadilla, Ortega, Hernando, & Bernal, 2012):

- **New community:** refers to the start-up of the recommender, when, although a catalog of items might exist, almost no users are present and the lack of user interaction makes it very hard to provide reliable recommendations.

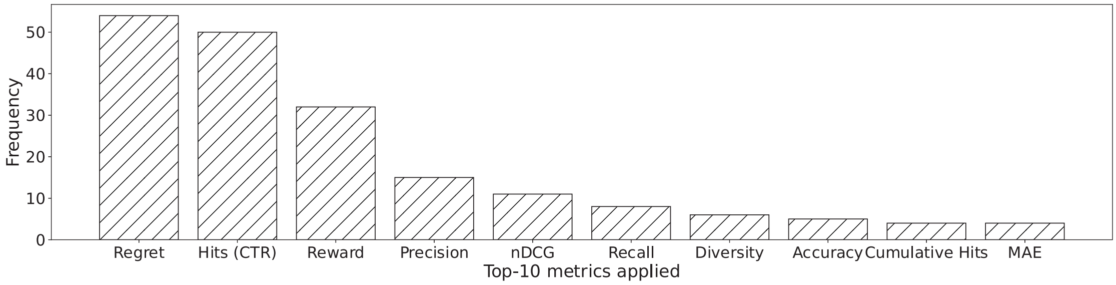


Fig. 9. Frequency of works in which is measured each of these evaluation metrics.

- **New item:** a new item is added to the system with some content information, but it has not have any review yet.
- **New user:** a new user make it registration and he/she has not provided any interaction yet (it is not possible to provide personalized recommendations).

In the literature, all of these cases have been explored by using MAB algorithms. In our SLR we identified that most of the works are related to the *new community* problem due to the evaluation policy applied. As the researchers have usually evaluated the learning quality of bandit algorithms, they simulate a start-up scenario without any information about user and items (Felício et al., 2017; Qiao, Yan, & Shen, 2018; Silva et al., 2020). On the other hand, other works have been proposed exclusively to face the *new item* and *new user* problems. The new item problem (or item cold-start) often is faced by a hybrid approach that combines the bandit algorithms with content-based (CB) concepts (Tang et al., 2014; Wang, Wang et al., 2017). Basically, these approaches use the items characteristics to correlate items and user's preferences when there is no any information about the items. In turn, we found two classes of bandit algorithms that are concerned with the *new user* problem (or user cold-start): (1) the Contextual MAB (CMAB); and (2) the hybrid approaches. The first one applies the contextual information about items or users (e.g., age, gender, contents, etc.) to allow users to explore the system's options (Nguyen & Lauw, 2014; Qin, Chen, & Zhu, 2014; Zeng et al., 2016). The second one proposes to apply a model selection to handle the first two stages of the user's experience, changing from one model to another when the system has enough data about the user (Felício et al., 2017; Gutowski, Amghar, Camp, & Chhel, 2019a; Gutowski et al., 2019b; Lacerda, 2017; Lacerda et al., 2015; Lu et al., 2018).

4.3.3. Privacy concerns

The growing concern about the privacy of users has affected several e-commerce systems and created many discussions in the recommendation field. The success of a recommendation is directly related to its personalization ability. Consequently, many algorithms try to collect as much data about the user as possible. In particular, this is even more evident in contextual algorithms (Nguyen & Lauw, 2014; Qin et al., 2014; Zeng et al., 2016). In order to improve the quality of the recommendations, such algorithms dynamically adapt to the users' interests and requirements based on the information collected by the recommendation agent. However, such information is often potentially confidential since it is associated with the users' particular interests in their previous interactions they may prefer to keep secret due to the content accessed, or even their addresses. In this sense, several privacy policies have been discussed over the years, leading many recommendation algorithms to change itself (Hannun, Knott, Sengupta, & van der Maaten, 2019; Malekzadeh, Athanasakis, Haddadi, & Livshits, 2019; Ren, Zhou, Liu, & Shroff, 2020). Even the multi-armed bandits' algorithms should be considered.

In the literature, we identified many proposals that can be applied in bandit algorithms. A first proposal is performing the recommendations in a particular device, like the users' own devices, in order to keep their personal information in their hands (Xin et al., 2015). However, a local recommendation does not present promise since the best personalization methods are often related to incorporate useful information from other users who have similar preferences. So, Malekzadeh et al. (2019) have proposed to apply a privacy model of *crowd-blending* (Gehrke, Hay, Lui, & Pass, 2012), where a user mixes himself with a number $l \geq 1$ of other users in a way that replacing him data with any other user does not change the results of the recommendation. So, their privacy is protected by this group of other users with the same interests. Another approach, proposed by Zhou et al. (2019), suggests storing the central recommendation model on a cloud server, only updating the learning parameters by each local recommendation agent. Then, all training data is kept on the local user agent and no updates are stored on the cloud server for privacy protection. Other works have discussed the privacy problem specifically for bandit algorithms by proposing a system that updates local agents by collecting feedback from other agents in a private manner (Hannun et al., 2019; Ren et al., 2020).

4.3.4. Explainable recommendations

Nowadays, there is a growing consensus that explanations have helped the user to better understand and interpret the rationale of the recommender system, thereby making it more trustworthy and engaging (Zhang & Chen, 2020). Famous platforms like Netflix, Amazon, and Spotify have labeled their recommendations by explanations such as 'because you watched it, you can like this'. Traditionally, their proposals usually follow one of two orthogonal dimensions: (1) the information source or display style of the explanations (e.g., textual sentence explanation, or visual explanation), which represents the human–computer interaction (HCI) perspective; and (2) model to generate such explanations, which represents the machine learning (ML) perspective. At the first one, various visualization techniques for explaining recommendations have been proposed, including interfaces with concentric circles (Kangasrääsiö, Glowacka, & Kaski, 2015) and pathways between columns (Bostandjiev, O'Donovan, & Höllerer, 2012). One of the most relevant is the taxonomy proposed by Friedrich and Zanker (2011) taking into account the style (e.g., collaborative, knowledge, utility or social explanation style) and paradigm (e.g. content-based, knowledge or collaborative based) and type of preference model. In turn, from the second perspective, other approaches have been proposed. Kouki, Schaffer, Pujara, O'Donovan, and Getoor (2017), for instance, propose a hybrid recommender system built on a probabilistic programming language and show that explanations improve the user experience of recommender systems.

In our SLR, we identify two works concerned with explaining the recommendations by applying a reinforcement learning algorithm. First, McInerney et al. (2018) proposed that users would respond to explanations differently and dynamically, and thus, a bandit-based approach for exploitation–exploration trade-off would help to find the best explanation orderings for each user. In particular, they proposed methods to jointly learn which explanations each user responds to, which are the best contents to recommend for each user, and how to balance exploration with exploitation to deal with uncertainty. This work shows that just as exploitation–exploration is beneficial to recommendation tasks, it is also beneficial to explanation tasks. Similarly, Wang, He, Feng, Nie and Chua (2018) proposed a model-agnostic reinforcement learning framework to generate sentence explanations for any recommendation model. In this design, the recommendation model to be explained is a part of the environment, while the agents are responsible for generating explanations and predicting the output ratings of the recommendation model based on the explanations. Thus, the agents learn to generate explanations with good explainability and presentation quality by optimizing the expected reward of their actions.

5. Future directions and research opportunities

As aforementioned, the applicability of MAB in the recommendation field is very recent and there still are a lot of research opportunities and improvements available for future works. In this section, we go beyond the discussions of what has been done and propose future directions to be concerned in future research. First, we highlight the main approaches that should be more studied or even improved according to the answers achieved by our first research question. Then, we open a new discussion in the field about the evaluation criteria based on the studies analyzed by answering our second research question. And, finally, based on our knowledge achieved by reading the works identified by our SLR, we present a brief discussion about how we can handle the challenges pointed out in the third question to make some improvements in the current literature.

5.1. Which algorithms should we design and explore in the future?

The first research question, answered by this work in Section 4.1, highlighted not only the high number of works published about MAB in the recommendation field, but also the main algorithms and strategies applied so far. We notice that among all MAB strategies, the UCB algorithm is the most chosen over the years to handle the exploration and exploitation dilemma due to two main reasons: (1st) the confidence bound measured by UCB was deeply studied in the literature and its effectiveness to measure the uncertainness about each arm has been already certified (Hariri et al., 2014; Zhou & Brunskill, 2016); and (2nd) UCB is more susceptible than other algorithms to combine new concepts and ideas due to its low complexity and the hyperparameter independence. Nowadays, UCB still is the main algorithm of MAB to be overcome in the recommendation field. Especially, current advances by combining UCB with classical Collaborative Filtering (CF) concepts (Wang, Zeng et al., 2018; Wu et al., 2016; Zhao et al., 2013) have achieved great performances. As shown in Fig. 7, approaches such as clustering, matrix factorization, and graph-based are the most usual concepts applied with UCB algorithms.

However, this usual choice for UCB does not mean that only this algorithm can be applied in the recommendation domain. We observed, for instance, that several works have explored the Thompson Sampling (TS) algorithm in the last five years (Kawale, Bui, Kveton, Thanh, & Chawla, 2015; Li, Karatzoglou et al., 2016; Zhu, Huang, & Xu, 2020a). We have usually noticed that researchers are trying to achieve the same TS performance from traditional reinforcement learning scenarios for this field. The main challenge is how to adapt the TS algorithm for the recommendation domain due to its complexity. Traditionally, the standard TS had performed poorly here due to the specific environment of recommendation systems that undergoes frequently. A recent work (Zhu et al., 2020a) has proposed to incorporate collaborative effects into TS by using the feedback of all users in the same dynamic clustering to estimate the expected reward in the current context. Despite they achieved a great convergence rate, the authors still do not contrast this new algorithm with traditional UCB approaches that have performed well. In general, this kind of advances remains open in the literature to be explored for future researchers.

Moreover, by analyzing the number of works published in each domain of the recommendation field in Fig. 4, we could notice that the major efforts are related to movies and news. As they are two relevant scenarios where the users are constantly interacting with the system during their session, it makes sense that they have been modeled by an online learning approach. However, after 15 years of studies in this topic, it is expected that research on other domains can be developed by the academic community. Famous companies from the music domain, such as Pandora and Spotify, have recently declared they interesting in MAB algorithms to improve the user's personalization (Semerci et al., 2019). In a recent talk on RecSys'19, Spotify's researchers showed that the quality of their recommendations depends on a MAB framework that balances exploration and exploitation to allow their model to adapt quickly to changes in users' preferences (Semerci et al., 2019). And, the same observation can be made for other scenarios like products and ads. In our opinion, the new researchers should address their efforts in this direction by evaluating the current algorithms in these scenarios and also by proposing new approaches.

5.2. Are the traditional evaluation criteria enough to measure the recommendation quality of bandit models?

By answering the second question proposed by our SLR in Section 4.2, we identified the most relevant metrics applied to measure the recommendation quality and highlighted the main evaluation methodologies defined for this field. In our discussions, we noticed that researchers have preferred to adapt traditional reward and regret metrics from the reinforcement learning scenario to this field. In general, these metrics are very similar to usual recommendation metrics applied to measure the system's prediction power, such as Hit Rate, MAE, MSE, and RMSE (Bobadilla et al., 2013). However, it is now widely agreed that accurate predictions are crucial but insufficient to deploy a good recommendation engine (Kunaver & Požrl, 2017; Vargas, 2014; Vargas & Castells, 2011). In many applications, people use a recommendation system for more than exact anticipation of their tastes. Users may also be interested in discovering new items, in rapidly exploring diverse items, in preserving their privacy, in the fast responses of the system, and many more properties of the interaction with the recommendation engine. Only prediction metrics are not enough to measure the actual quality of a recommendation. In our opinion, user satisfaction metrics, such as novelty, diversity, serendipity, unexpectedness, and others should be more used in this field by the researchers.

Moreover, we also believe that the current criteria of evaluation by trials (i.e., concerned into estimate the rating for each user–item pair at each interaction) is not ideal to measure the system's quality. By choosing only one user–item pair to be evaluated at each system iteration, the current works are especially focused on the learning ability of their model. However, the recommendation task is not only about the system. In its basic definition, a good recommendation system is the one that helps **users** to find the most **relevant items** according to their **needs** and **preferences**. Thus, we think that MAB should also walk in the same direction as the recommendation community by exploring some methodologies focused on the users' interactions. A promising strategy to measure the recommendation quality would be regarded on each user's interaction with the system individually (at least when a decision is taken online). In a real-world scenario, there is not only one user interacting with the system per time (as simulated by trial methodologies). Moreover, the system's response to the users' interactions should be fast because users will not wait for many trials to receive huge updates in the items' popularity or in the best products.

Especially, the lack of rigorous and well-defined evaluation criteria to measure the quality of the current MAB algorithms is a big problem for the community. Nowadays, it is not easy to replicate a simple bandit algorithm because we do not have a specific evaluation protocol where the system was applied. Most of the time, the researchers have to replicate the algorithm and the evaluation criteria defined for that specific scenario to make sure they have implemented everything correctly. In our opinion, a framework developed for the recommendation scenario, concerning the user's needs and preferences, will be a huge contribution to both academy and industry.

5.3. How can we handle the challenges studied in the literature?

Answering the third research question of our SLR in Section 4.3, we discuss five distinct recommendation challenges that have received a lot of attention in the current years. They are undoubtedly the most relevant challenges to be addressed actually in the MAB scope. The question that remains open is definitely how we can handle these problems to make some improvements in the current literature. As much as the literature has many approaches to partially solve each of them, we have noticed that some solutions are actually guiding the system to other problems. Facing data sparsity, for instance, we figured out that current works have proposed one of two possible solutions: (1) combine the bandit algorithm with clustering or factorization based techniques to make predictions based on a group of interests; or (2) introduce contextual bandit algorithms to reduce the lack of information about users and items. In the first one, the main techniques often applied are non-scalable to make online recommendations by clustering the users' preferences in real-time. And the second one, in addition to the scalability problem that a lot of contexts can represent, we also have to concern about some privacy issues when the context is crawled. Until now, the most applied solution for the sparsity and scalability problems is related to the first one by performing updates in clusters (or latent factors) during the learning stage to avoid the retrain process (Nguyen

& Lauw, 2014; Qiao et al., 2018; Wang, Zeng et al., 2018; Yang et al., 2020). However, even this proposal must be improved to also fetch collaborative feedbacks by considering the opinions of other users in a scalable way.

Similarly, despite being widely studied in the MAB scope, the cold-start problem still is a challenge for the current algorithms. In particular, it happens because most of the existing works are often misleading the problem definition when they affirm to handle the cold-start. To the best of our knowledge, facing the new community problem is not similar to face new users or new items problem individually (Adomavicius & Tuzhilin, 2005; Bobadilla et al., 2012). When the algorithm is proposed to face the new community problem, it is simulating a new system where there is not any information about users and items (e.g., when the Netflix, Spotify, or other platform was announced). It is not similar to face the new users (items) problem that usually happens in real-world scenarios, where new users (items) are joining the system that already has previous data. In the first scenario, the system does not even know which item is the most popular or the best-rated because all the community is new. On the other hand, in the new user (item) scenario, the most popular items are already defined because there is an entire community interacting with the system. Then, when a MAB algorithm handles the new community problem does not mean it can face the new users (items). The existing bias of a community can improve the recommendation once we already know the items with more probability to be rated, or even undermine the recommendation quality by favoring only these items from others (Bobadilla et al., 2013). Unfortunately, we realized the main MAB works have neglected the impact of the new user (or new item) cold-start problems in their MAB implementations. It is necessary to explore in more detail what happens with the current MAB algorithms when a new user (or a new item) is added to the community. By analyzing the recommendation quality from the user perspective (i.e., by applying our suggestion of Section 5.2), most MAB algorithms will generate bad recommendations for users in the first interactions.

Furthermore, as highlighted before, there are only a few works concerned about the privacy and explainability of new bandit algorithms (Hannun et al., 2019; Malekzadeh et al., 2019). They are relevant topics in MAB to the industry and they probably will receive even more attention. Especially, despite the relevance of explainability in the recommendation field, this work does not find any work concerned into explain the recommendations provided by a bandit algorithm. In our opinion, this lack of works may be related to how MAB algorithms are applied. As they are usually combined with traditional RSs techniques, the researchers may be assuming that the existing approaches are enough to explain the recommendations. However, we have not found any work to prove this argument. So, we also suggest new researchers explore this topic and improve community knowledge in this direction.

6. Conclusion

In this work we have presented a systematic literature review of Multi-Armed Bandits in the recommendation field to shed light upon their applicability and open challenges. By inspecting 1327 articles published from the last twenty years (2000–2020), we identified 230 works as the most relevant studies about MAB in the field. These articles were read in detail and analyzed to fill a specific data extraction form. This form guides this work to achieve three main goals: (1) it consolidates an updated picture of the main research conducted in this area so far; (2) it highlights the most used concepts and methods, their core characteristics, and their main limitations; and (3) it evaluates the applicability of MAB-based recommendation approaches in some traditional RSs' challenges, such as data sparsity, scalability, cold-start, privacy, and explainability.

The discussions proposed by our work indicate that several advances were already achieved by the existing works. Until now, these advances have attracted the attention of important conferences and journals, such as RecSys, WWW, NIPS, ICML, CIKM, and others. In general, our analyzes indicate that most existing work focuses on UCB or Thompson Sampling algorithms combined with concepts of collaborative filtering, such as Matrix Factorization, clustering approaches, and others. Especially, by inspecting these works in detail, we also presented a standard implementation for each traditional MAB algorithm (ε-Greedy, UCB, and TS) when they are combined with matrix factorization concepts. Moreover, the analyzes of this work also identify the main methodologies and metrics applied to simulate the online environment. In the inspected works, we observe a simple abstraction of the traditional evaluation criteria of reinforcement learning scenarios, where the system only concerns about the model's learning by measuring the reward or regret assigned. It, unfortunately, means that current methodologies have only concerned with the prediction power of the recommendation — an outdated concept in the current recommendation literature that is more worried about the user's satisfaction or engagement.

In general, although many interesting works have been elaborated in this research field, there are still many challenges that need to be tackled. First, we consider that some efforts should be done to overcome the current approaches of UCB and to better adapt the TS algorithm for the recommendation domain. Second, there is a need to define rigorous evaluation criteria to measure the quality of the current algorithms here, in the recommendation field. In our opinion, the traditional evaluation approach of classical scenarios of reinforcement learning is not compatible with this field, where the usual goal is to help users with their needs. Third, based on our knowledge achieved by reading the works identified by our SLR, we believe that the traditional challenges of the recommendation domain should be more studied. Most of them, like the cold-start problem, can directly affect the performance of current bandit algorithms. In short, we hope this work can contribute as an interesting starting point with many ideas for this new topic by helping to further shape relevant research on Multi-Armed Bandits.

CRediT authorship contribution statement

Nícollas Silva: Conceptualization, Methodology, Papers reading, Filling the data extract form, Validation, Formal analysis, Writing – review & editing. **Heitor Werneck:** Papers reading, Filling the data extract form, Validation, Plot graphics, Write tables , Writing – review & editing. **Thiago Silva:** Papers search, Papers reading, Filling the data extract form, Validation, Writing – review & editing. **Adriano C.M. Pereira:** Supervision, Methodology, Validation, Formal analysis, Writing – review & editing, Project administration. **Leonardo Rocha:** Supervision, Methodology, Validation, Formal analysis, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was partially supported by CNPq, Brazil, CAPES, Brazil, FINEP, Brazil, Fapemig, Brazil, and INWEB, Brazil.

References

Adomavicius, G., & Tuzhilin, A. (2005). 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*, *17*(6), 734–749.

Aharon, M., Kagian, A., Kaplan, Y., Nissim, R., & Somekh, O. (2015). Serving ads to" Yahoo Answers" occasional visitors. In *Proceedings of the 24th international conference on world wide web* (pp. 1257–1262).

Auer, P. (2002). Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, *3*(Nov), 397–422.

Auer, P., Cesa-Bianchi, N., & Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, *47*(2–3), 235–256.

Bagaria, V., Kamath, G., Ntranos, V., Zhang, M., & Tse, D. (2018). Medoids in almost-linear time via multi-armed bandits. In *International conference on artificial intelligence and statistics* (pp. 500–509). PMLR.

Balakrishnan, A., Bouneffouf, D., Mattei, N., & Rossi, F. (2019). Using multi-armed bandits to learn ethical priorities for online AI systems. *IBM Journal of Research and Development*, *63*(4/5), 1.

Barraza-Urbina, A., Koutrika, G., d'Aquin, M., & Hayes, C. (2018). BEARS: Towards an evaluation framework for bandit-based interactive recommender systems. In *REVEAL 18, October 6-7, 2018*. Vancouver, Canada: NUI Galway.

Basu, S., Sen, R., Sanghavi, S., & Shakkottai, S. (2019). Blocking bandits. In *Advances in neural information processing systems* (pp. 4784–4793). URL: http://papers.nips.cc/paper/8725-blocking-bandits.

Bernardi, L., Estevez, P., Eidis, M., & Osama, E. (2020). Recommending accommodation filters with online learning. In J. Vinagre, A. M. Jorge, M. Al-Ghossein, & A. Bifet (Eds.), *CEUR workshop proceedings*: *vol. 2715*, *Proceedings of the 3rd workshop on online recommender systems and user modeling co-located with the 14th ACM conference on recommender systems (RecSys 2020), Virtual Event, September 25, 2020*. CEUR-WS.org, URL: http://ceur-ws.org/Vol-2715/paper3.pdf.

Bobadilla, J., Ortega, F., Hernando, A., & Bernal, J. (2012). A collaborative filtering approach to mitigate the new user cold start problem. *Knowledge-Based Systems*, *26*, 225–238.

Bobadilla, J., Ortega, F., Hernando, A., & Gutiérrez, A. (2013). Recommender systems survey. *Knowledge-Based Systems*.

Bostandjiev, S., O'Donovan, J., & Höllerer, T. (2012). TasteWeights: a visual interactive hybrid recommender system. In *Proceedings of the sixth ACM conference on recommender systems* (pp. 35–42).

Bouneffouf, D. (2014). Freshness-aware Thompson sampling. In *International conference on neural information processing* (pp. 373–380). Springer, URL: https://link.springer.com/chapter/10.1007/978-3-319-12643-2_46.

Bouneffouf, D. (2016). Contextual bandit algorithm for risk-aware recommender systems. In *2016 IEEE congress on evolutionary computation (CEC)* (pp. 4667–4674). IEEE.

Bouneffouf, D., Bouzeghoub, A., & Gançarski, A. L. (2012a). A contextual-bandit algorithm for mobile context-aware recommender system. In *International conference on neural information processing* (pp. 324–331). Springer.

Bouneffouf, D., Bouzeghoub, A., & Gançarski, A. L. (2012b). Hybrid-ε-greedy for mobile context-aware recommender system. In *Pacific-Asia conference on knowledge discovery and data mining* (pp. 468–479). Springer.

Bouneffouf, D., Bouzeghoub, A., & Gançarski, A. (2012c). Following the user's interests in mobile context-aware recommender systems: The hybrid-e-greedy algorithm. (pp. 657–662). http://dx.doi.org/10.1109/WAINA.2012.200.

Bouneffouf, D., & Claeys, E. (2016). Learning exploration for contextual bandit. In *AutoML ICML 2019: 6th ICML workshop on automated machine learning*.

Bouneffouf, D., Laroche, R., Urvoy, T., Féraud, R., & Allesiardo, R. (2014). Contextual bandit for active learning: Active Thompson sampling. In *International conference on neural information processing* (pp. 405–412). Springer.

Bouneffouf, D., Rish, I., Cecchi, G. A., & Féraud, R. (2017). Context attentive bandits: Contextual bandit with restricted context. arXiv preprint arXiv:1705.03821.

Bresler, G., Chen, G., & Shah, D. (2014). A latent source model for online collaborative filtering. *Advances in Neural Information Processing Systems*, *4*.

Brodén, B., Hammar, M., Nilsson, B. J., & Paraschakis, D. (2018). Ensemble recommendations via Thompson sampling: an experimental study within e-commerce. In *23rd international conference on intelligent user interfaces* (pp. 19–29). URL: https://dl.acm.org/doi/abs/10.1145/3172944.3172967.

Brodén, B., Hammar, M., Nilsson, B. J., & Paraschakis, D. (2019). A bandit-based ensemble framework for exploration/exploitation of diverse recommendation components: An experimental study within e-commerce. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *10*(1), 1–32.

Cañamares, R., Redondo, M., & Castells, P. (2019). Multi-armed recommender system bandit ensembles. In *Proceedings of the 13th ACM conference on recommender systems* (pp. 432–436). URL: https://dl.acm.org/doi/abs/10.1145/3298689.3346984.

Çano, E., & Morisio, M. (2017). Hybrid recommender systems: A systematic literature review. *Intelligent Data Analysis*, *21*(6), 1487–1524.

Cao, Y., Wen, Z., Kveton, B., & Xie, Y. (2019). Nearly optimal adaptive procedure with change detection for piecewise-stationary bandit. In *The 22nd international conference on artificial intelligence and statistics* (pp. 418–427). URL: http://proceedings.mlr.press/v89/cao19a.html.

Caron, S., & Bhagat, S. (2013). 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* (pp. 1–9). URL: https://dl.acm.org/doi/abs/10.1145/2501025.2501029.

Castells, P., Hurley, N. J., & Vargas, S. (2015). Novelty and diversity in recommender systems. In *Recommender systems handbook* (pp. 881–918). Springer.

Castells, P., Vargas, S., & Wang, J. (2011). Novelty and diversity metrics for recommender systems: Choice, discovery and relevance. In *Proceedings of international workshop on diversity in document retrieval (DDR)*.

Celis, L. E., Kapoor, S., Salehi, F., & Vishnoi, N. (2019). Controlling polarization in personalization: An algorithmic framework. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 160–169).

Cesa-Bianchi, N., Gentile, C., & Zappella, G. (2013). A gang of bandits. In *Advances in neural information processing systems* (pp. 737–745).

Chapelle, O., & Li, L. (2011). An empirical evaluation of Thompson sampling. In *Advances in neural information processing systems* (pp. 2249–2257).

Chatterji, N., Muthukumar, V., & Bartlett, P. (2020). Osom: A simultaneously optimal algorithm for multi-armed and linear contextual bandits. In *International conference on artificial intelligence and statistics* (pp. 1844–1854).

Chen, L., Krause, A., & Karbasi, A. (2017). Interactive submodular bandit. In *Advances in neural information processing systems* (pp. 141–152).

Chen, M., & Liu, P. (2017). Performance evaluation of recommender systems. *International Journal of Performability Engineering*, *13*(8).

Chen, L., Xu, J., & Lu, Z. (2018). Contextual combinatorial multi-armed bandits with volatile arms and submodular reward. In *Advances in neural information processing systems* (pp. 3247–3256).

Chi, C. M., Lin, H. T., & Ing, C. K. (2019). Online clustering of bandits with high-dimensional sparse relevant user features. In *Proceedings of the 22nd international conference on artificial intelligence and statistics (AISTATS)*. PMLR.

Christakopoulou, K., & Banerjee, A. (2018). Learning to interact with users: A collaborative-bandit approach. In *Proceedings of the 2018 SIAM international conference on data mining* (pp. 612–620). SIAM, arXiv:https://www-users.cs.umn.edu/~baner029/papers/18/sdm18-bandit.pdf. URL: https://epubs.siam.org/doi/abs/10.1137/1.9781611975321.69.

Christakopoulou, K., Radlinski, F., & Hofmann, K. (2016). Towards conversational recommender systems. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 815–824).

Crammer, K., & Gentile, C. (2011). Multiclass classification with bandit feedback using adaptive regularization. *Machine Learning*, *90*, 273–280. http://dx.doi.org/10.1007/s10994-012-5321-8.

Cremonesi, P., Garzotto, F., Negro, S., Papadopoulos, A. V., & Turrin, R. (2011). Looking for "good" recommendations: A comparative evaluation of recommender systems. In *IFIP conference on human-computer interaction* (pp. 152–168). Springer.

Duchi, J. (2017). *CS229 supplemental lecture notes Hoeffding's inequality*. Stanford University.

Dumitrascu, B., Feng, K., & Engelhardt, B. (2018). PG-TS: Improved thompson sampling for logistic contextual bandits. *Advances in Neural Information Processing Systems*, *31*, 4624–4633.

edw (2020). Selecting multiple web adverts: A contextual multi-armed bandit with state uncertainty. *Journal of the Operational Research Society*, *71*(1), 100–116, URL: https://www.tandfonline.com/doi/abs/10.1080/01605682.2018.1546650.

Eide, S., & Zhou, N. (2018). Deep neural network marketplace recommenders in online experiments. In *Proceedings of the 12th ACM conference on recommender systems* (pp. 387–391).

Felício, C., Paixão, K., Barcelos, C., & Preux, P. (2017). A multi-armed bandit model selection for cold-start user recommendation. In *Proceedings of the 25th conference on user modeling, adaptation and personalization*.

Friedrich, G., & Zanker, M. (2011). A taxonomy for generating explanations in recommender systems. *AI Magazine*, *32*(3), 90–98.

Gehrke, J., Hay, M., Lui, E., & Pass, R. (2012). Crowd-blending privacy. In *Annual cryptology conference* (pp. 479–496). Springer.

Gentile, C., Li, S., Kar, P., Karatzoglou, A., Zappella, G., & Etrue, E. (2017). On context-dependent clustering of bandits. In *International conference on machine learning* (pp. 1253–1262). PMLR.

Gentile, C., Li, S., & Zappella, G. (2014). Online clustering of bandits. In *International conference on machine learning* (pp. 757–765).

Geyik, S. C., Dialani, V., Meng, M., & Smith, R. (2018). In-session personalization for talent search. In *Proceedings of the 27th ACM international conference on information and knowledge management* (pp. 2107–2115).

Gomes, L., Almeida, C., & Vale, Z. (2020). Recommendation of workplaces in a coworking building: A cyber-physical approach supported by a context-aware multi-agent system. *Sensors*, *20*(12), 3597.

Goswami, A., Zhai, C., & Mohapatra, P. (2019). Learning to diversify for E-commerce search with multi-armed bandit. In *ECOMSIGIR*. URL: http://ceur-ws.org/Vol-2410/paper18.pdf.

Guillou, F., Gaudel, R., & Preux, P. (2016). Scalable explore-exploit collaborative filtering. In *Pacific Asia conference on information systems (PACIS)*. Association For Information System.

Gupta, S., Balaji, B., & Luo, R. (2020). CPR: Collaborative pairwise ranking for online list recommendations. In J. Vinagre, A. M. Jorge, M. Al-Ghossein, & A. Bifet (Eds.), *CEUR workshop proceedings*: *vol. 2715*, *Proceedings of the 3rd workshop on online recommender systems and user modeling co-located with the 14th ACM conference on recommender systems (RecSys 2020), Virtual Event, September 25, 2020*. CEUR-WS.org, URL: http://ceur-ws.org/Vol-2715/paper9.pdf.

Gutowski, N., Amghar, T., Camp, O., & Chhel, F. (2019a). Global versus individual accuracy in contextual multi-armed bandit. In *Proceedings of the 34th ACM/SIGAPP symposium on applied computing* (pp. 1647–1654).

Gutowski, N., Amghar, T., Camp, O., & Chhel, F. (2019b). Gorthaur: A portfolio approach for dynamic selection of multi-armed bandit algorithms for recommendation. In *2019 IEEE 31st international conference on tools with artificial intelligence (ICTAI)* (pp. 1164–1171). IEEE.

Gutowski, N., Amghar, T., Camp, O., & Hammoudi, S. (2017). A framework for context-aware service recommendation for mobile users: A focus on mobility in smart cities. In *From data to decision*.

Gutowski, N., Camp, O., Amghar, T., & Chhel, F. (2019c). Using individual accuracy to create context for non-contextual multi-armed bandit problems. In *2019 IEEE-RIVF international conference on computing and communication technologies (RIVF)* (pp. 1–6). IEEE.

Gutowski, N., Camp, O., Chhel, F., Amghar, T., & Albers, P. (2019). Improving bandit-based recommendations with spatial context reasoning: An online evaluation. In *2019 IEEE 31st international conference on tools with artificial intelligence (ICTAI)* (pp. 1366–1373). IEEE.

Hannun, A., Knott, B., Sengupta, S., & van der Maaten, L. (2019). Privacy-preserving contextual bandits. arXiv preprint arXiv:1910.05299.

Hao, B., Yadkori, Y. A., Wen, Z., & Cheng, G. (2019). Bootstrapping upper confidence bound. In *Advances in neural information processing systems* (pp. 12123–12133). URL: http://papers.nips.cc/paper/9382-bootstrapping-upper-confidence-bound.

Hariri, N., Mobasher, B., & Burke, R. (2014). Context adaptation in interactive recommender systems. In *Proceedings of the 8th ACM conference on recommender systems* (pp. 41–48).

Hariri, N., Mobasher, B., & Burke, R. (2015). Adapting to user preference changes in interactive recommendation. In *Twenty-fourth international joint conference on artificial intelligence*.

Heckel, R., & Ramchandran, K. (2017). The sample complexity of online one-class collaborative filtering. In *Proceedings of the 34th international conference on machine learning JMLR*.

Hsieh, C.-C., Neufeld, J., King, T., & Cho, J. (2015). Efficient approximate Thompson sampling for search query recommendation. In *Proceedings of the 30th annual ACM symposium on applied computing* (pp. 740–746).

Immorlica, N., Mao, J., Slivkins, A., & Wu, Z. S. (2019). Bayesian exploration with heterogeneous agents. In *The world wide web conference* (pp. 751–761).

Jagerman, R., Markov, I., & de Rijke, M. (2019). When people change their mind: Off-policy evaluation in non-stationary recommendation environments. In *Proceedings of the twelfth ACM international conference on web search and data mining* (pp. 447–455).

Jannach, D., Zanker, M., Felfernig, A., & Friedrich, G. (2010). *Recommender systems: an introduction*. Cambridge University Press.

Jedor, M., Perchet, V., & Louedec, J. (2019). Categorized bandits. In *Advances in neural information processing systems* (pp. 14422–14432).

(2011). Personalized pricing recommender system: Multi-stage epsilon-greedy approach. In *Proceedings of the 2nd international workshop on information heterogeneity and fusion in recommender systems* (pp. 57–64). URL: https://dl.acm.org/doi/abs/10.1145/2039320.2039329.

Kangasrääsiö, A., Glowacka, D., & Kaski, S. (2015). Improving controllability and predictability of interactive recommendation interfaces for exploratory search. In *Proceedings of the 20th international conference on intelligent user interfaces* (pp. 247–251).

Katariya, S., Kveton, B., Wen, Z., & Potluru, V. K. (2019). Conservative exploration using interleaving. In *The 22nd international conference on artificial intelligence and statistics* (pp. 954–963). URL: http://proceedings.mlr.press/v89/katariya19a.html.

Kawale, J., Bui, H. H., Kveton, B., Thanh, L. T., & Chawla, S. (2015). Efficient Thompson sampling for online matrix-factorization recommendation. *Advances in Neural Information Processing Systems*, *28*, 1297–1305.

Keele, S., et al. (2007). *Guidelines for performing systematic literature reviews in software engineering: Technical Report Technical report, Ver. 2.3*, EBSE Technical Report. EBSE.

Khenissi, S., Mariem, B., & Nasraoui, O. (2020). Theoretical modeling of the iterative properties of user discovery in a collaborative filtering recommender system. In *Fourteenth ACM conference on recommender systems* (pp. 348–357).

Kong, W., Brunskill, E., & Valiant, G. (2020). Sublinear optimal policy value estimation in contextual bandits. In *International conference on artificial intelligence and statistics* (pp. 4377–4387). PMLR.

Kouki, P., Schaffer, J., Pujara, J., O'Donovan, J., & Getoor, L. (2017). User preferences for hybrid explanations. In *Proceedings of the eleventh ACM conference on recommender systems* (pp. 84–88).

Krohn-Grimberghe, A., Nanopoulos, A., & Schmidt-Thieme, L. (2010). A novel multidimensional framework for evaluating recommender systems. In *LWA* (pp. 113–120).

Kunaver, M., & Požrl, T. (2017). Diversity in recommender systems–A survey. *Knowledge-Based Systems*, *123*, 154–162.

Lacerda, A. (2017). Multi-objective ranked bandits for recommender systems. *Neurocomputing*, *246*, 12–24, URL: https://www.sciencedirect.com/science/article/pii/S092523121730228X.

Lacerda, A., Veloso, A., & Ziviani, N. (2015). Adding value to daily-deals recommendation: Multi-armed bandits to match customers and deals. In *2015 Brazilian conference on intelligent systems (BRACIS)* (pp. 216–221). IEEE.

Li, H. (2011). A recommendation system in cognitive radio networks with random data traffic. *IEEE Transactions on Vehicular Technology*, *60*(4), 1352–1364.

Li, L., Chu, W., Langford, J., & Schapire, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on world wide web* (pp. 661–670).

Li, L., Chu, W., Langford, J., & Wang, X. (2011). Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of the fourth ACM international conference on web search and data mining* (pp. 297–306).

Li, M., Jiang, W., & Li, K. (2017). When and what music will you listen to? Fine-grained time-aware music recommendation. In *2017 IEEE international symposium on parallel and distributed processing with applications and 2017 IEEE international conference on ubiquitous computing and communications (ISPA/IUCC)* (pp. 1091–1098). IEEE.

Li, S., Karatzoglou, A., & Gentile, C. (2016). Collaborative filtering bandits. In *Proceedings of the 39th international ACM SIGIR conference on research and development in information retrieval* (pp. 539–548).

Li, S., Wang, B., Zhang, S., & Chen, W. (2016). Contextual combinatorial cascading bandits. In *ICML, Vol. 16* (pp. 1245–1253). URL: http://www.jmlr.org/proceedings/papers/v48/lif16-supp.pdf.

Liang, Y., Loni, B., & Larson, M. A. (2017). CLEF NewsREEL 2017: Contextual bandit news recommendation. In *CLEF (Working notes)*. arXiv:https://pure.tudelft.nl/portal/files/35743499/35743393.pdf. URL: https://pure.tudelft.nl/portal/files/35743499/35743393.pdf.

Liu, C., Cai, Q., & Zhang, Y. (2017). Multi-armed bandit mechanism with private histories. In *Proceedings of the 16th conference on autonomous agents and multiagent systems* (pp. 1607–1609).

Liu, W., Li, S., & Zhang, S. (2018). Contextual dependent click bandit algorithm for web recommendation. In *International computing and combinatorics conference* (pp. 39–50). Springer, arXiv:https://shuaili8.github.io/Publications/cocoon2018.pdf. URL: https://link.springer.com/chapter/10.1007/978-3-319-94776-1_4.

Liu, B., Wei, Y., Zhang, Y., Yan, Z., & Yang, Q. (2018). Transferable contextual bandit for cross-domain recommendation. In *AAAI*. URL: https://openreview.net/forum?id=r1-g8CxdWB&noteId=r1-g8CxdWB.

Louëdec, J., Chevalier, M., Mothe, J., Garivier, A., & Gerchinovitz, S. (2015). A multiple-play bandit algorithm applied to recommender systems. In *FLAIRS conference* (pp. 67–72).

Lu, X., Wen, Z., & Kveton, B. (2018). Efficient online recommendation via low-rank ensemble sampling. In *Proceedings of the 12th ACM conference on recommender systems* (pp. 460–464). URL: https://dl.acm.org/doi/abs/10.1145/3240323.3240408.

Mahajan, D. K., Rastogi, R., Tiwari, C., & Mitra, A. (2012). Logucb: an explore-exploit algorithm for comments recommendation. In *Proceedings of the 21st ACM international conference on information and knowledge management* (pp. 6–15). URL: https://dl.acm.org/doi/abs/10.1145/2396761.2396767.

Malekzadeh, M., Athanasakis, D., Haddadi, H., & Livshits, B. (2019). Privacy-preserving bandits.

Manickam, I., Lan, A. S., & Baraniuk, R. G. (2017). Contextual multi-armed bandit algorithms for personalized learning action selection. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 6344–6348). IEEE.

Maniu, S., Ioannidis, S., & Cautis, B. (2020). Bandits under the influence (extended version). arXiv preprint arXiv:2009.10135.

Martín, M., Jiménez-Martín, A., & Mateos, A. (2019). A numerical analysis of allocation strategies for the multi-armed bandit problem under delayed rewards conditions in digital campaign management. *Neurocomputing*, *363*, 99–113.

Matikainen, P., Furlong, P., Sukthankar, R., & Hebert, M. (2013). Multi-armed recommendation bandits for selecting state machine policies for robotic systems. In *Proceedings - IEEE international conference on robotics and automation* (pp. 4545–4551). http://dx.doi.org/10.1109/ICRA.2013.6631223.

May, B. C., Korda, N., Lee, A., & Leslie, D. S. (2012). Optimistic Bayesian sampling in contextual-bandit problems. *Journal of Machine Learning Research*, *13*(1), 2069–2106, arXiv:http://www.jmlr.org/papers/volume13/may12a/may12a.pdf. URL: https://dl.acm.org/doi/abs/10.5555/2503308.2343711.

McInerney, J., Lacker, B., Hansen, S., Higley, K., Bouchard, H., Gruson, A., et al. (2018). Explore, exploit, and explain: personalizing explainable recommendations with bandits. In *Proceedings of the 12th ACM conference on recommender systems* (pp. 31–39). URL: https://dl.acm.org/doi/abs/10.1145/3240323.3240354.

Mehrotra, R., Xue, N., & Lalmas, M. (2020). 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* (pp. 3224–3233).

Mishra, N., & Thakurta, A. (2014). Private stochastic multi-arm bandits: From theory to practice. In *ICML workshop on learning, security, and privacy*.

Mukherjee, S., Kveton, B., & Rao, A. B. (2019). Latent ranked bandits. In *ICML 2019 workshop RL4RealLife*.

Nguyen, H. T., & Kofod-Petersen, A. (2014). Using multi-armed bandit to solve cold-start problems in recommender systems at telco. In *Mining intelligence and knowledge exploration* (pp. 21–30). Springer.

Nguyen, T. T., & Lauw, H. W. (2014). Dynamic clustering of contextual multi-armed bandits. In *Proceedings of the 23rd ACM international conference on conference on information and knowledge management* (pp. 1959–1962).

Nguyen, M. N., Pham, C., Son, J., & Hong, C. S. (2016). Online learning-based clustering approach for news recommendation systems. In *2016 18TH Asia-Pacific network operations and management symposium (APNOMS)* (pp. 1–4). IEEE.

Park, D. H., Kim, H. K., Choi, I. Y., & Kim, J. K. (2012). A literature review and classification of recommender systems research. *Expert Systems with Applications*, *39*(11), 10059–10072.

Pathak, B., Garfinkel, R., Gopal, R. D., Venkatesan, R., & Yin, F. (2010). Empirical analysis of the impact of recommender systems on sales. *Journal of Management Information Systems*, *27*(2), 159–188.

Peng, Y., Xie, M., Liu, J., Meng, X., Li, N., Yang, C., et al. (2019). A practical semi-parametric contextual bandit.. In *IJCAI* (pp. 3246–3252). URL: https://www.ijcai.org/Proceedings/2019/0450.pdf.

Qiao, R., Yan, S., & Shen, B. (2018). A reinforcement learning solution to cold-start problem in software crowdsourcing recommendations. In *2018 IEEE international conference on progress in informatics and computing (PIC)* (pp. 8–14). IEEE.

Qin, L., Chen, S., & Zhu, X. (2014). Contextual combinatorial bandit and its application on diversified online recommendation. In *Proceedings of the 2014 SIAM international conference on data mining* (pp. 461–469). SIAM, URL: https://epubs.siam.org/doi/abs/10.1137/1.9781611973440.53.

Rahman, M., & Oh, J. C. (2015a). Fast online learning to recommend a diverse set from big data. In *International conference on industrial, engineering and other applications of applied intelligent systems* (pp. 361–370). Springer.

Rahman, M., & Oh, J. C. (2015b). Parallel and synchronized UCB2 for online recommendation systems. In *2015 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology (WI-IAT), Vol. 1* (pp. 413–416). IEEE.

Rahman, M., & Oh, J. C. (2018). Graph bandit for diverse user coverage in online recommendation. *Applied Intelligence: The International Journal of Artificial Intelligence, Neural Networks, and Complex Problem-Solving Technologies*, *48*(8), 1979–1995.

Rao, D. (2020). Contextual bandits for adapting to changing user preferences over time. arXiv preprint arXiv:2009.10073.

Ravi, A. N., Poduval, P., & Moharir, S. (2020). Unreliable multi-armed bandits: A novel approach to recommendation systems. In *2020 international conference on COMmunication Systems & NETworkS (COMSNETS)* (pp. 650–653). IEEE.

Ren, W., Zhou, X., Liu, J., & Shroff, N. B. (2020). Multi-armed bandits with local differential privacy. arXiv preprint arXiv:2007.03121.

Santana, M. R., Melo, L. C., Camargo, F. H., Brandão, B., Soares, A., Oliveira, R. M., et al. (2020). Contextual meta-bandit for recommender systems selection. In *Fourteenth ACM conference on recommender systems* (pp. 444–449).

Sanz-Cruzado, J., Castells, P., & López, E. (2019). A simple multi-armed nearest-neighbor bandit for interactive recommendation. In *Proceedings of the 13th ACM conference on recommender systems* (pp. 358–362).

Saritaç, A. O., & Tekin, C. (2017). Combinatorial multi-armed bandit problem with probabilistically triggered arms: A case with bounded regret. In *2017 IEEE global conference on signal and information processing (GlobalSIP)* (pp. 111–115). IEEE.

Sato, M., Nagatani, K., & Tahara, T. (2017). Exploring an optimal online model for new job recommendation: Solution for recsys challenge 2017. In *Proceedings of the recommender systems challenge 2017* (pp. 1–5).

Semerci, O., Gruson, A., Edwards, C., Lacker, B., Gibson, C., & Radosavljevic, V. (2019). Homepage personalization at spotify. In *Proceedings of the 13th ACM conference on recommender systems* (pp. 527–527).

Shani, G., & Gunawardana, A. (2011). Evaluating recommendation systems. In *Recommender systems handbook* (pp. 257–297). Springer.

Shapira, B., Ricci, F., Kantor, P. B., & Rokach, L. (2011). *Recommender systems handbook*. Springer.

Silva, T., Silva, N., Werneck, H., Pereira, A. C., & Rocha, L. (2020). The impact of first recommendations based on exploration or exploitation approaches in recommender systems' learning. In *Proceedings of the Brazilian symposium on multimedia and the web* (pp. 173–180).

Song, L., Fragouli, C., & Shah, D. (2018). Recommender systems over wireless: Challenges and opportunities. In *2018 IEEE information theory workshop (ITW)* (pp. 1–5). IEEE.

Song, L., Fragouli, C., & Shah, D. (2019). Interactions between learning and broadcasting in wireless recommendation systems. In *2019 IEEE international symposium on information theory (ISIT)* (pp. 2549–2553). IEEE.

Song, L., Tekin, C., & Van Der Schaar, M. (2014a). Clustering based online learning in recommender systems: A bandit approach. In *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 4528–4532). IEEE.

Song, L., Tekin, C., & Van Der Schaar, M. (2014b). Online learning in large-scale contextual recommender systems. *IEEE Transactions on Services Computing*, *9*(3), 433–445, URL: https://ieeexplore.ieee.org/abstract/document/6940318/.

Sui, Y., Gotovos, A., Burdick, J., & Krause, A. (2015). Safe exploration for optimization with Gaussian processes. In *International conference on machine learning* (pp. 997–1005). PMLR, URL: http://proceedings.mlr.press/v37/sui15.html.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT Press.

Takemori, S., Sato, M., Sonoda, T., Singh, J., & Ohkuma, T. (2020). Submodular bandit problem under multiple constraints. arXiv preprint arXiv:2006.00661.

Tang, L., Jiang, Y., Li, L., & Li, T. (2014). Ensemble contextual bandits for personalized recommendation. In *Proceedings of the 8th ACM conference on recommender systems* (pp. 73–80).

Tang, L., Jiang, Y., Li, L., Zeng, C., & Li, T. (2015). Personalized recommendation via parameter-free contextual bandits. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval* (pp. 323–332).

Tavakol, M., & Brefeld, U. (2017). A unified contextual bandit framework for long-and short-term recommendations. In *Joint European conference on machine learning and knowledge discovery in databases* (pp. 269–284). Springer.

Tekin, C., & Turğay, E. (2018). Multi-objective contextual multi-armed bandit with a dominant objective. *IEEE Transactions on Signal Processing*, *66*(14), 3799–3813.

Tekin, C., & Van Der Schaar, M. (2015). RELEAF: An algorithm for learning and exploiting relevance. *IEEE Journal of Selected Topics in Signal Processing*, *9*(4), 716–727.

Tekin, C., Zhang, S., & van der Schaar, M. (2014). Distributed online learning in social recommender systems. *IEEE Journal of Selected Topics in Signal Processing*, *8*(4), 638–652.

Teo, C. H., Nassif, H., Hill, D., Srinivasan, S., Goodman, M., Mohan, V., et al. (2016). Adaptive, personalized diversity for visual discovery. In *Proceedings of the 10th ACM conference on recommender systems* (pp. 35–38). URL: https://dl.acm.org/doi/abs/10.1145/2959100.2959171.

Theocharous, G., Vlassis, N., & Wen, Z. (2017). An interactive points of interest guidance system. In *Proceedings of the 22nd international conference on intelligent user interfaces companion* (pp. 49–52).

Tracà, S., Rudin, C., & Yan, W. (2020). Reducing exploration of dying arms in mortal bandits. In *Uncertainty in artificial intelligence* (pp. 156–163). PMLR.

Tripathi, A., Ashwin, T., & Guddeti, R. M. R. (2018). A reinforcement learning and recurrent neural network based dynamic user modeling system. In *2018 IEEE 18th international conference on advanced learning technologies (ICALT)* (pp. 411–415). IEEE.

Vargas, S. (2011). New approaches to diversity and novelty in recommender systems. In *Fourth BCS-IRSG symposium on future directions in information access (FDIA 2011) 4* (pp. 8–13).

Vargas, S. (2014). Novelty and diversity enhancement and evaluation in recommender systems and information retrieval. In *Proceedings of the 37th international ACM SIGIR conference on research & development in information retrieval* (pp. 1281–1281).

Vargas, S., & Castells, P. (2011). Rank and relevance in novelty and diversity metrics for recommender systems. In *Proceedings of the fifth ACM conference on recommender systems* (pp. 109–116).

Vorobev, A., Lefortier, D., Gusev, G., & Serdyukov, P. (2015). Gathering additional feedback on search results by multi-armed bandits with respect to production ranking. (pp. 1177–1187). http://dx.doi.org/10.1145/2736277.2741104.

Wang, X., He, X., Feng, F., Nie, L., & Chua, T.-S. (2018). Tem: Tree-enhanced embedding model for explainable recommendation. In *Proceedings of the 2018 world wide web conference* (pp. 1543–1552).

Wang, X., Hoi, S. C., Liu, C., & Ester, M. (2017). Interactive social recommendation. In *Proceedings of the 2017 ACM on conference on information and knowledge management* (pp. 357–366).

Wang, Q., Li, T., Iyengar, S., Shwartz, L., & Grabarnik, G. Y. (2018). Online it ticket automation recommendation using hierarchical multi-armed bandit algorithms. In *Proceedings of the 2018 SIAM international conference on data mining* (pp. 657–665). SIAM.

Wang, P., Liu, K., Jiang, L., Li, X., & Fu, Y. (2020). Incremental mobile user profiling: Reinforcement learning with spatial knowledge graph for modeling event streams. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 853–861).

Wang, Y., Ouyang, H., Wang, C., Chen, J., Asamov, T., & Chang, Y. (2017). Efficient ordered combinatorial semi-bandits for whole-page recommendation. In *AAAI* (pp. 2746–2753). URL: http://www.yichang-cs.com/yahoo/AAAI17_SemiBandits.pdf.

Wang, X., Wang, Y., Hsu, D., & Wang, Y. (2014). Exploration in interactive personalized music recommendation: a reinforcement learning approach. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, *11*(1), 1–22.

Wang, L., Wang, C., Wang, K., & He, X. (2017). Biucb: A contextual bandit algorithm for cold-start and diversified recommendation. In *2017 IEEE international conference on big knowledge (ICBK)* (pp. 248–253). IEEE.

Wang, H., Wu, Q., & Wang, H. (2016). Learning hidden features for contextual bandits. In *Proceedings of the 25th ACM international on conference on information and knowledge management* (pp. 1633–1642).

Wang, H., Wu, Q., & Wang, H. (2017). Factorization bandits for interactive recommendation. In *Thirty-first AAAI conference on artificial intelligence*.

Wang, Q., Zeng, C., Zhou, W., Li, T., Iyengar, S. S., Shwartz, L., et al. (2018). Online interactive collaborative filtering using multi-armed bandit with dependent arms. *IEEE Transactions on Knowledge and Data Engineering*, *31*(8), 1569–1580.

Wang, H., Zhao, Q., Wu, Q., Chopra, S., Khaitan, A., & Wang, H. (2020). Global and local differential privacy for collaborative bandits. In *Fourteenth ACM conference on recommender systems* (pp. 150–159).

Warlop, R., Lazaric, A., & Mary, J. (2018). Fighting boredom in recommender systems with linear reinforcement learning. *Advances in Neural Information Processing Systems*, *31*, 1757–1768.

Wen, Y., Wang, F., Wu, R., Liu, J., & Cao, B. (2020). Improving the novelty of retail commodity recommendations using multiarmed bandit and gradient boosting decision tree. *Concurrency Computations: Practice and Experience*, Article e5703.

Wu, X., Cetintas, S., Kong, D., Lu, M., Yang, J., & Chawla, N. (2020). Learning from cross-modal behavior dynamics with graph-regularized neural contextual bandit. In *Proceedings of the web conference 2020* (pp. 995–1005).

Wu, Q., Iyer, N., & Wang, H. (2018). Learning contextual bandits in a non-stationary environment. In *The 41st international ACM SIGIR conference on research & development in information retrieval* (pp. 495–504). URL: https://dl.acm.org/doi/abs/10.1145/3209978.3210051.

Wu, Q., Wang, H., Gu, Q., & Wang, H. (2016). Contextual bandits in a collaborative environment. In *Proceedings of the 39th international ACM SIGIR conference on development in information retrieval*.

Wu, Q., Wang, H., Hong, L., & Shi, Y. (2017). Returning is believing: Optimizing long-term user engagement in recommender systems. In *Proceedings of the 2017 ACM on conference on information and knowledge management* (pp. 1927–1936). arXiv:https://par.nsf.gov/servlets/purl/10066038. URL: https://dl.acm.org/doi/abs/10.1145/3132847.3133025.

Wu, Q., Wang, H., Li, Y., & Wang, H. (2019). Dynamic ensemble of contextual bandits to satisfy users' changing interests. In *WWW '19: The world wide web conference* (pp. 2080–2090). http://dx.doi.org/10.1145/3308558.3313727.

Wu, Q., Zhang, H., Gao, X., He, P., Weng, P., Gao, H., et al. (2019). Dual graph attention networks for deep latent representation of multifaceted social effects in recommender systems. In *The world wide web conference* (pp. 2091–2102). URL: https://dl.acm.org/doi/abs/10.1145/3308558.3313442.

Wu, M., Zhu, Y., Yu, Q., Rajendra, B., Zhao, Y., Aghdaie, N., et al. (2019). A recommender system for heterogeneous and time sensitive environment. In *Proceedings of the 13th ACM conference on recommender systems* (pp. 210–218).

Xin, Y., et al. (2015). *Challenges in recommender systems: scalability, privacy, and structured recommendations* (Ph.D. thesis), Massachusetts Institute of Technology.

Xu, X., Dong, F., Li, Y., He, S., & Li, X. (2020). Contextual-bandit based personalized recommendation with time-varying user interests. In *AAAI* (pp. 6518–6525).

Xu, X., Vakili, S., Zhao, Q., & Swami, A. (2017). Online learning with side information. In *MILCOM 2017-2017 IEEE military communications conference (MILCOM)* (pp. 303–308). IEEE.

Yan, Y., Liu, Z., Zhao, M., Guo, W., Yan, W. P., & Bao, Y. (2018). A practical deep online ranking system in e-commerce recommendation. In *Joint European conference on machine learning and knowledge discovery in databases* (pp. 186–201). Springer.

Yang, M., Li, Q., Qin, Z., & Ye, J. (2020). Hierarchical adaptive contextual bandits for resource constraint based recommendation. In *Proceedings of the web conference 2020* (pp. 292–302).

Yang, L., Liu, B., Lin, L., Xia, F., Chen, K., & Yang, Q. (2020). Exploring clustering of bandits for online recommendation system. In *Fourteenth ACM conference on recommender systems* (pp. 120–129).

Yang, K., & Toni, L. (2018). Graph-based recommendation system. In *2018 IEEE global conference on signal and information processing (GlobalSIP)* (pp. 798–802). IEEE.

Yu, B., Fang, M., & Tao, D. (2016). Linear submodular bandits with a knapsack constraint. In *Proceedings of the thirtieth AAAI conference on artificial intelligence* (pp. 1380–1386).

Yu, T., Mengshoel, O., Meroux, D., & Jiang, Z. (2019). *Machine learning with decision trees and multi-armed bandits: An interactive vehicle recommender system*: Technical Report, SAE Technical Paper, URL: https://www.sae.org/publications/technical-papers/content/2019-01-1079/.

Yu, T., Shen, Y., & Jin, H. (2019). A visual dialog augmented interactive recommender system. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 157–165).

Yu, T., Shen, Y., & Jin, H. (2020). Towards hands-free visual dialog interactive recommendation. In *Proceedings of the AAAI conference on artificial intelligence, Vol. 34* (pp. 1137–1144).

Yue, Y., & Guestrin, C. (2011). Linear submodular bandits and their application to diversified retrieval. In *Advances in neural information processing systems* (pp. 2483–2491).

Zeng, C., Wang, Q., Mokhtari, S., & Li, T. (2016). 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* (pp. 2025–2034).

Zhang, Y., & Chen, X. (2020). Explainable recommendation: A survey and new perspectives. http://dx.doi.org/10.1561/9781680836592.

Zhang, H., Sun, Y., Qiu, J., Zhu, C., Zhao, J., & Wang, H. (2020). A music recommendation system based on reinforcement learning. *Design Engineering*, 331–342.

Zhang, X., Xie, H., Li, H., & C.S. Lui, J. (2020). Conversational contextual bandit: Algorithm and application. In *Proceedings of the web conference 2020* (pp. 662–672).

Zhang, X., Zhou, Q., He, T., & Liang, B. (2018). Con-CNAME: A contextual multi-armed bandit algorithm for personalized recommendations. In *International conference on artificial neural networks* (pp. 326–336). Springer.

Zhao, T., & King, I. (2016). Locality-sensitive linear bandit model for online social recommendation. In *International conference on neural information processing* (pp. 80–90). Springer.

Zhao, C., Watanabe, K., Yang, B., & Hirate, Y. (2018). Fast converging multi-armed bandit optimization using probabilistic graphical model. In *Pacific-Asia conference on knowledge discovery and data mining* (pp. 115–127). Springer.

Zhao, X., Xia, L., Tang, J., & Yin, D. (2019). Deep reinforcement learning for search, recommendation, and online advertising: a survey. *ACM SIGWEB Newsletter*, (Spring), 1–15.

Zhao, C., Yang, B., & Hirate, Y. (2019). A reward optimization model for decision-making under budget constraint. *Journal of Information Processing*, *27*, 190–200, URL: https://www.jstage.jst.go.jp/article/ipsjjip/27/0/27_190/_article/-char/ja/.

Zhao, X., Zhang, W., & Wang, J. (2013). Interactive collaborative filtering. In *Proceedings of the 22nd ACM international conference on information & knowledge management* (pp. 1411–1420).

Zhong, S., Ying, W., Chen, X., & Fu, Q. (2020). An adaptive similarity-measuring-based CMAB model for recommendation system. *IEEE Access*, *8*, 42550–42561.

Zhou, L., & Brunskill, E. (2016). Latent contextual bandits and their application to personalized recommendations for new users. In *Proceedings of the 25th international joint conference on artificial intelligence (IJCAI)*.

Zhou, S., Dai, X., Chen, H., Zhang, W., Ren, K., Tang, R., et al. (2020). 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* (pp. 179–188).

Zhou, C., Jin, Y., Wang, X., & Zhang, Y. (2020). Conversational music recommendation based on bandits. In *2020 IEEE international conference on knowledge graph (ICKG)* (pp. 41–48). IEEE.

Zhou, P., Wang, K., Guo, L., Gong, S., & Zheng, B. (2019). A privacy-preserving distributed contextual federated online learning framework with big data support in social recommender systems. *IEEE Transactions on Knowledge and Data Engineering*.

Zhu, Z., Huang, L., & Xu, H. (2020a). Collaborative thompson sampling. *Mobile Networks and Applications*, 1–13.

Zhu, Z., Huang, L., & Xu, H. (2020b). Self-accelerated thompson sampling with near-optimal regret upper bound. *Neurocomputing*.

Zhu, Y., Lin, J., He, S., Wang, B., Guan, Z., Liu, H., et al. (2019). Addressing the item cold-start problem by attribute-driven active learning. *IEEE Transactions on Knowledge and Data Engineering*, *32*(4), 631–644.

Zong, S., Ni, H., Sung, K., Ke, N. R., Wen, Z., & Kveton, B. (2016). Cascading bandits for large-scale recommendation problems. In *32nd Conf. on uncertainty in artificial intelligence (UAI), 2016*.