

Online Experimentation for Information Retrieval

Katja Hofmann^(✉)

Microsoft Research, Cambridge, UK
katja.hofmann@microsoft.com

Abstract. Online experimentation for information retrieval (IR) focuses on insights that can be gained from user interactions with IR systems, such as web search engines. The most common form of online experimentation, A/B testing, is widely used in practice, and has helped sustain continuous improvement of the current generation of these systems.

As online experimentation is taking a more and more central role in IR research and practice, new techniques are being developed to address, e.g., questions regarding the scale and fidelity of experiments in online settings. This paper gives an overview of the currently available tools. This includes techniques that are already in wide use, such as A/B testing and interleaved comparisons, as well as techniques that have been developed more recently, such as bandit approaches for online learning to rank.

This paper summarizes and connects the wide range of techniques and insights that have been developed in this field to date. It concludes with an outlook on open questions and directions for ongoing and future research.

Keywords: Online evaluation · A/B testing · Contextual bandits · Dueling bandits · Interleaved comparison · Online learning to rank · Counterfactual analysis · Experiment design

1 Introduction

Online experimentation for information retrieval (IR) refers to experiments that rely on natural user interactions. For example, an *online evaluation* experiment might compare alternative search interfaces, or alternative methods for ranking search results (often referred to as *rankers*). Such controlled experiments allow researchers or system developers to gain a better understanding of how to support the searchers' goals, or to test models of information seeking behavior. Extending the controlled experiment scenario, *online learning* approaches can automate the experimentation process to efficiently search a large space of IR solutions.

Many examples and success stories of online evaluation have been published in previous years. For example, Kohavi et al. [44] show how AB testing was used to identify a search widget for MSN Real Estate that would maximize revenue.

The result of the experiment described there led to a 10% increase in revenue, illustrating that online experiments can lead to high real-world impact that often cannot be predicted even by domain experts.

The techniques that have been developed for online experimentation for IR complement more traditional experimentation techniques. Test-collection based IR experimentation [63, 75] can efficiently compare search solutions at various levels of abstraction, but require expensive manual annotations. For example, they allow experimenters to focus on an isolated concept like topical relevance, while abstracting from individual differences between searchers. In contrast, online experimentation techniques have been developed to evaluate and tune systems to directly optimize systems' online performance. As a result, they reflect expectations and behavior of real users and can adapt to their preferences.

While the initial focus of online experimentation was primarily on improving the performance of a given system, they are not only applicable to system development, but can also provide new insights from a research perspective. Online experimentation is particularly closely related to research in interactive IR [38]. The focus of interactive IR research has led to valuable insights in experimental design, and has been particularly informative in testing hypotheses and developing theory, e.g., of search behavior. This more theoretical view can benefit online experimentation, showing a path beyond the optimization of individual system performance. Correspondingly, techniques developed for online experimentation can add to the tool set available for interactive IR research, in particular where unobtrusive measurement is required for naturalistic studies.

The focus of online experimentation on studying natural user interactions in realistic settings results in a unique set of challenges. First, exploring solutions of unknown performance creates the risk of hurting the user experience. At the same time, feedback on these solutions can only be obtained by trying them out. This results in a trade-off between exploration of new solutions, and exploitation of good solutions known at a given point in time. Second, in a natural setting we often cannot control many sources of variance (e.g., search goal) that would be controlled in a more traditional experiment, such as a lab study. This can result in high variance, but this problem is typically alleviated by large sample sizes that can be collected, at least in frequently-used systems. Finally, online experiments in natural settings can offer a very narrow windows of observation. Instead of e.g., recordings and follow-up interviews that may be collected in a lab study, and that can help interpret results, we now have to rely exclusively on behavior traces (e.g., clicks, page views) that result from users' natural interactions. This limited bandwidth of observation can be particularly problematic when unobserved confounding variables affect measured outcomes (e.g., effects of search task or user characteristics). Therefore it is particularly important to carefully design the experiment, and especially the measurements designed to evaluate or compare solutions.

In the remainder of this overview paper, we outline the techniques that have been developed to address these challenges. First, we further motivate the need for controlled experiments, and introduce online evaluation using the example

of A/B testing (Sect. 2). The next two sections focus on measurement, first in the form of estimating online metrics from exploration data (Sect. 3), then in the form of paired comparisons enabled by interleaved comparisons (Sect. 5). After discussing questions related to online evaluation, we turn to the question of how these evaluation techniques can be used to automatically optimize system performance, for example when many system configurations are feasible. We turn to online learning, where we first introduce a common problem formulation – bandits (Sect. 6). Finally, we focus on learning from relative feedback (Sect. 7). Sections 6 and 7 build on the earlier sections, in that many of the learning approaches utilize the previously introduced online evaluation methods to infer feedback for learning.

2 Controlled Experiments

This section outlines the role that controlled experiments play in identifying causal relationships. We establish the connection between these concepts and IR, and discuss examples of controlled experiments, from small-scale lab studies to web-scale experiments.

Scientific discovery can take many forms. Following the three-fold distinction of Babbie [5], exploratory studies are valuable for identifying phenomena of interest and formulating hypotheses. Descriptive studies describe phenomena and their observed relationships. Finally, explanatory studies aim to uncover *causal relationships*, that explain the mechanisms that lead to the observed phenomena and relationships.

Identifying causal relationships is particularly valuable, because they are the most robust [54, 55]. By explaining the mechanisms of why events happen, they allow us to make predictions under changing conditions. These insights are crucial for predicting the effects of actions. In everyday life, knowing causal relationships and consequences of our actions lets us make informed decisions on how to achieve our goals. For developers of an interactive system, identifying causal relationships enables data-driven decisions on how the system should interact with the user.

In our everyday life, we are used to thinking in terms of cause and effect, and many causal relationships are obvious to us. Following an example of Babbie [5], when we observe a correlation between ice cream consumption and death by drowning in lakes, we would not conclude that one causes the other. Rather, we know that there is a common cause, temperature or season, that affects both.

In many cases, causal structures are far less obvious, especially when systems are complex, and potential causes cannot be observed. In particular, observational data alone is not enough to draw any conclusions on causal relationships [55]. Observing a correlation (also called association) between events could result from infinitely many possible causal relationships. This is illustrated in the path diagram in Fig. 1. Returning to the example above, observing a correlation between ice cream consumption and deaths by drowning does not, by itself, allow us to infer the causal structure that explains these events. Only with

The observation:



X and Y are correlated.

May result from infinitely many causal relationships, such as:



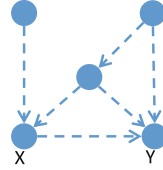
X causes Y.



Y causes X.



X and Y have a common cause that affects both.



More complex causal structures.

Fig. 1. Path diagram of an example observation (X and Y are correlated), and possible causal structures that may explain the observation. Possible correlation between two variables is denoted by bi-directional arrows. Hypothesized causal relations are denoted by directed arrows from cause to effect.

additional information can we reason that the common cause explanation is the most likely. Conversely, the absence of correlation in observational data does not exclude the possibility of a causal relationship.

For a concrete example from information retrieval, let us consider two studies of searchers' click behavior in web search. Granka et al. [22] report on an observational study, which measures the time searchers spend looking at search result summaries (document titles and snippets) per rank, and the number of clicks per rank. They find that searchers spend more time examining higher-ranked documents, and that they click on higher-ranked documents more often than on lower-ranked documents. The findings describe searcher behavior, but note that it does not allow us to draw conclusions about causal relationships. For example, it is possible that examination and clicks are caused by document rank, or by some other factor, such as the attractiveness of the snippets (e.g., if more attractive snippets are ranked higher). In IR, we are often interested in document relevance, but again, observational data alone does not allow us to draw conclusions on whether relevance may be causally related to the observed behavior.

The most reliable method for identifying causal relationships are controlled experiments [55]. To explain observed correlations between click behavior, search result rank, and document relevance, Guan and Cutrell [23] conducted a controlled experiment in a lab study on search behavior. They manipulated the search result pages shown to study participants to include a single relevant document, and randomly selected the rank at which this target document was shown. They found that, when the target was ranked lower in the result list, participants were often not able to find it, and were likely to click on less relevant higher-ranked documents. The results suggest that both rank and document relevance affect searchers' click behavior. Based on this insight, numerous click models have been proposed that model document relevance using causal assumptions about rank, and observed user interactions (e.g., [14, 18]).

Conducting a controlled experiment means that, instead of observing naturally occurring values of all variables, we specifically set the value of the hypothesized causal variable (e.g., X in Fig. 1). In doing so, we break the causal chain that may carry associations between this variable and the hypothesized effect (e.g., Y). The key to a successful intervention is that the assignment of values to X is done at random. This ensures that the decision to assign the chosen value is independent of any other variables that could carry information between X and Y . Any remaining relationship between X and Y then has to reflect the strength of the causal relationship between the two variables.

Designing controlled experiments may be difficult or impossible in some settings, e.g., when studying the economy of a country. In some of these cases, it may be possible to infer causal structure from initial causal assumptions in combination (i.e., non-experimental data that was observed outside of a controlled experiment) [55]. In the present discussion we focus on settings where controlled experiments are possible. Combining the insights that result from controlled experiments with causal inference mechanisms can result in an even more powerful discovery process.

Luckily, many interactive systems are well suited for experimental control. For example, if we are interested in whether a redesign of a website affects user engagement, we can conduct a controlled experiment (also called “A/B test” or “bucket test” in the context of web-scale studies) [43–45, 72]. We can do this by deciding for each new incoming user at random whether they are directed to the original version (often called control) or the redesigned version (often called treatment). We measure the target quantity we are interested in (e.g., number of click, or time spent on the page). When comparing the measurements for control and treatment group, statistical significance testing is used to establish whether any observed differences are likely to result from random noise [11]. If statistically significant differences are observed, these can be attributed solely to the differences between the two versions of the website, because our random assignment to control and treatment group blocked any other possible causal effects. Once a causal relationship has been established, it can guide decisions on how to change the current system.

Controlled experiments can be conducted in interactive systems of any scale, ranging from small-scale laboratory experiments to crowdsourced experiments with hundreds or thousands of crowdworkers, to experiments with millions of users for large web systems. Each of these affords a different level of control. In lab studies, where it is typically only feasible to work with a small number of participants, it is often useful to work with a complex experimental design that investigate several variables at once (see [38] for a detailed discussion of experiment design, especially in interactive IR studies). For example, in an eye-tracking study of user interactions with query auto-completion (QAC), we investigated the effect of QAC quality on 10 variables that captured user behavior, while controlling for effects of search task and user differences [31]. Crowdsourcing environments can afford an interesting balance between scale and control. For example, Kazai et al. [37] studied the effect of quality-control mechanisms

and page ordering on the quality of annotations in a relevance labeling task with hundreds of workers. At web scale, experimental designs are often less complex, to keep results interpretable. For example, Bendersky et al. [7] examined the effects of retrieval-based video recommendation strategies on users’ viewing behavior in a month-long experiment that involved millions of users.

The methodology for running controlled experiments on the web, in the form of A/B tests, has been refined over recent years. Kohavi et al. [43] gives a detailed account of the key aspects that need to be considered, such as deciding on a metric, implementing the split in control and treatment group, and estimating the required sample size. An important challenge is the question bandwidth available for running experiments. Often, online experiments run for several days or weeks, and running only one experiment at a time would lead to very slow progress (especially considering that typically few of the tested changes improve system performance or lead to significant insights) [45]. This can be improved by running several mutually independent experiments in parallel [43, 72]. Improving the sensitivity of online controlled experiments is an area of active research [19].

This section motivated the need for controlled experiments as a basis for evaluating interactive systems. The most prominent methodology for running these experiments online, A/B testing, was briefly introduced. In the next section, we discuss extensions of the controlled experiment setup that allow the use of exploration for large-scale offline evaluation.

3 Offline Estimates of Online Performance

In the previous section we discussed how controlled experiments allow us to assess effects of system changes metrics on their users. In this section, we introduce methods that allow system comparisons using so-called exploration data. Recall that the key requirement for controlled experiments is that the assignment of users to control and treatment conditions has to be independent of any other factors that may carry information between the hypothesized cause and effect, and that randomization is a reliable way of blocking any such interactions. The same principle is exploited in the methods introduced here.

The first approach we discussed is called exploration scavenging [48].¹ The question raised there is whether a data set that was collected under an arbitrary data collection policy² π_D can be used to evaluate alternative policies π_A (i.e., other configurations of the system). The problem is formulated as the task of obtaining an unbiased and consistent estimate of the online performance of π_A

¹ The terminology comes from the area of reinforcement learning, a type of machine learning in which an intelligent agent (e.g., an interactive IR system) learns from interactions with its environment (e.g., users) by trying out actions and observing rewards. This is a natural model for learning in interactive IR, and is discussed in more detail in Sect. 6.

² A policy defines a distribution over system actions, often conditioned on additional information, such as the history of previous interactions, or information about context, such as a query posed by the user.

in terms of a target metric (e.g., clickthrough rate – CTR) offline, i.e., without running an actual experiment and using previously collected data. Langford et al. [48] show that this is not the case generally, but that there are cases where exploration data can be used to obtain unbiased estimates of the online performance of alternative policies. In particular, this is the case when π_D selects actions independently of the information used by π_A , and when it selects all actions that are available to π_A sufficiently often.

Li et al. [49, 50] propose and analyze a specific exploration scheme, and show that it allows very accurate prediction of online performance. Their approach relies on an exploration policy that selects actions during data collection uniformly at random. They show that this data can be used to obtain unbiased offline estimates of online performance. The method uses rejection sampling, where an observed sample is accepted to contribute to the estimate if it matches the action that would have been selected by the system under evaluation, and rejected otherwise. The effectiveness of the method is demonstrated in the context of a news recommendation application, where the task is to learn how to select the news article to display the most prominently to maximize user engagement (in terms of CTR). Recently, Li et al. [51] demonstrated an application of unbiased estimation from exploration data to optimize components of a commercial search engines (here: speller) in a large parameter space. They also propose non-uniform sampling during exploration, and show that very accurate estimates of online performance can be obtained.

The key benefit of the proposed offline evaluation techniques is that they allow infinitely many system comparisons once a set of exploration data has been collected. When exploration was done uniformly at random, meaning that it is independent of any information that the system might use, we can evaluate any system, including those where decisions are based on user attributes, or interaction history. This creates a powerful set up for testing effects of these factors on user interactions, and can be used to optimize system performance directly in terms of online metrics.

The methods proposed in [48–51] are powerful when the number of actions available to a system are relatively small compared to the sample size, e.g., when selecting from a small set of news articles, or from a small set of ads to be placed on the result page for a relatively frequent query. The amount of data required to obtain accurate (low-variance) estimates grows linearly in the number of available actions, making the approaches infeasible for large action spaces. An alternative is proposed in [8], where exploration is not in terms of the set of available actions, but instead in terms of the parameter space of the system.

Bottou et al. [8] propose a counterfactual reasoning approach. In it, exploration is achieved by changing system decisions from being deterministic to following some distribution (e.g., instead of using a fixed setting for a given parameter, use a continuous distribution over that parameter, from which the actual value for each impression is sampled during data collection). Given exploration data collected this way, counterfactual reasoning can be used to answer

“what if” questions of the form “*What would have happened if we had used a different system configuration?*”. Answers are obtained using importance sampling, where observed samples are reweighted by the ratio of their probability during data collection, and in the system under evaluation. Finally, Bottou et al. [8] propose a learning approach that utilizes counterfactual reasoning to compute the direction of parameter updates.

The counterfactual reasoning approach extends the principle of controlled experiments to settings where it is impossible to split a user population into control and treatment. The experimental unit is the impression, and every user experiences various parameter settings at various times. This allows the method to be applied to complex system, which Bottou et al. [8] demonstrate in the example of an online advertising marketplace.

The methods for offline evaluation that have been discussed in this section are inspired by methods from reinforcement learning, where off-policy evaluation is used to greatly speed up learning [56]. They enable methods for learning in interactive IR that directly maximize online performance (Sect. 6), but can naturally be used for system evaluation, or experimentation to test IR models that go beyond optimizing the performance of a specific application. In addition to the possibilities these methods open up for individual researchers or teams, the idea of using exploration data for unbiased evaluation may open up a path to sharing (anonymized) data in a future form of test-collection based IR evaluation.

So far, we have assumed that our goal is to evaluate systems in terms of an arbitrary online evaluation metric. Insights into what to measure, and proposed online evaluation metrics, are discussed in the next section (Sect. 4). The idea of using exploration data for offline evaluation is extended to within-subject experiments in the context of interleaved comparisons, discussed in Sect. 5.

4 Online Metrics

While implicit feedback has been used for IR evaluation for a long time [39], ubiquitous access to the web and web search engines have emphasized the need for reliable and interpretable online metrics. Consequently, research efforts in this direction have dramatically intensified in recent years and much progress has been made. Because this is such a large and active research area, the overview here only mentions a number of selected approaches that illustrate certain trends and developments. More detailed considerations regarding measurement in interactive experiments can be found in [40].

The specific choice of target metric very much depends on the specific application, and the goals of the experiment. For commercial applications, revenue, the number of purchases, or the value of purchases per buyer [43]. Similar metrics can be considered for recommendation systems and advertising platforms. As a general measure of user engagement, Dupret and Lalmas [21] recently proposed modeling absence time, i.e., how long a user waits before returning to a website. Crucially, the target metric should be decided on before the experiment is run. Other considerations include the variance of the metric and the expected difference between systems, as these affect the size of the sample required to detect

statistically significant differences between system (the power of a controlled experiment) [43].

In search, it is notoriously difficult to identify a single reliable online metrics. For example, changes in the number of clicks per query might mean that the user needs more clicks to find what they are looking for, or that several highly relevant pages are shown and keep the user engaged. An increase in abandonment rate could indicate that searchers give up in frustration, or that they can easily find the answer to their question directly on the search result page. Especially user clicks were shown to be affected by biases, e.g., due to result presentation [27, 82], and to vary substantially across search tasks and users [64]. Consequently, these and similar absolute evaluation metrics have been found to exhibit high variance, and caution has to be taken in interpreting their results [61].

Many recently proposed metrics take a more long-term or holistic view on measuring search engine quality. Joachims et al. [41] used per-query type models of dwell time to capture user satisfaction with search results, and personalized models of user satisfaction are explored in [24]. Song et al. [69] analyzed the long-term behavior of metrics, and showed that users may initially compensate for changes in search engine quality. Absence time as a measure of search effectiveness was considered in [12].

The interpretation of user signals as relative feedback has been proposed as an alternative to high-variance absolute metrics. Joachims et al. [35] show that the interpretation of clicks as relative preferences between documents, using so-called click-skip heuristics, can lead to accurate relative judgments. A proposed aggregation of these rules into a result page-level metric is PSkip [76]. Based on the construction of a controlled experiment, FairPairs infers the relative preference between documents from their relative CTR [59].

A difficulty with per-document relative metrics can be the large amount of exploration required for obtaining accurate estimates. This problem is avoided by interleaved comparison methods, which aggregate interactions with documents into a ranking level comparison. This can be thought of in similar terms as the exploration strategies over specific actions as opposed to exploration in terms of system parameters discussed in the previous section. Interleaved comparison methods are discussed in the next section.

5 Interleaved Comparisons

Interleaved comparisons have been developed to provide unbiased, relative comparisons of ranked lists [61]. In comparison to A/B tests, which run controlled experiments between users (each user is either in the control or in the treatment condition), interleaving experiments can be thought of as a within-subject experiment, where each user is presented with results that combine two competing rankings. To avoid introducing bias in such a setting, the interleaved (combined) result lists presented to users need to be constructed in a way that is fair to both rankers in expectation, and it has to be ensured that users cannot distinguish between the results contributed by either ranker.

The most widely-known interleaved comparison methods is Team-Draft interleaving [15, 57, 61]. We briefly describe the general principle of interleaving using this method. In Team-Draft interleaving, interleaved result lists are constructed in a way that is designed to ensure that each original ranker contributes the same number of its documents at a given rank to a given rank of the target list in expectation over impressions. This is done as follows. To fill the first two ranks in the interleaved list, a coin-flip determines which rankers first contributes a document. This ranker deterministically choses its highest document that is not yet part of the interleaved list. Then the competing ranker contributes its highest-ranked document. The process continues until a result list of the desired length has been constructed. During interleaving, the system keeps track of which ranker contributed which document. The constructed interleaved result list is then shown to the user, and user clicks (or, potentially, other interactions) on the presented documents are recorded. The observed clicks are then interpreted as preference indications for one of the rankers. Only the clicks on results contributed by a ranker are counted in its favor. Aggregating over multiple impressions results in an estimate of how much a ranker would be preferred over its competitor.

Team-Draft interleaving constructs a controlled within-subject experiment to compare between two rankers. This setup is extended by Probabilistic Interleave [26, 30]. Probabilistic Interleave is based on the idea of generating interleaved result lists from probability distributions over documents. These distributions are based on the rankings to be compared, in order to maintain the fairness of the interleaving. Interleaving outcomes can be computed by marginalizing over possible ways in which the observed interleaved result list could be generated. The result is a highly sensitive comparison method in which the magnitude of assigned click weights reflects the magnitude of ranking differences between the original rankings.

Crucially, probabilistic interleave defines an exploration policy, similar to those discussed in Sect. 3. This means that the collected data can be used to obtain unbiased estimates of online metrics, in this case of online interleaved comparison outcomes [28]. This results in a flexible online/offline evaluation setup, where interleaved comparisons can be performed online, and the observed data can be used as exploration data for further comparisons. Conversely, exploration data that was not collected using interleaving, but covers the same action space, can be used to obtain interleaved comparison outcomes for rankers in that same space. In Sect. 7 we show how the resulting method can be used to learn very efficiently from interleaving feedback.

Several extensions of interleaving have been devised. Optimized interleaving [58] considers the construction of a distribution over interleaved lists as a constrained optimization problem designed to obtain accurate comparisons between known rankers from as few samples as possible. Vertical-aware interleaving shows that the interleaved comparison approach can be extended to settings where the linear ranking assumption is violated, e.g., in the presence of vertical search results [16, 17]. Most recently, multileave was proposed to efficiently compare sets of rankers without having to perform all pairwise comparisons [67].

Interleaved comparison methods are particularly interesting for online evaluation because they allow within-subject experiments that result in particularly low variance. This provides highly sensitive comparisons with up to two orders of magnitude smaller sample sizes than those that would be required for comparable A/B tests [57].

6 Online Learning for Information Retrieval

Up to now we have focused on the use of controlled experiments for online evaluation. Given one or more systems, online evaluation techniques assess their absolute or relative online performance. However, in an online system, it is often not necessary to accurately determine the online performance of each candidate system. Rather, we are interested in identifying the best performing system as quickly as possible. When we need to choose from a small fixed set of systems, the earlier we know which one performs best, the sooner we can stop exploring the alternatives. In this setting, online learning can avoid over-exploring sub-optimal systems, leading to better online performance while learning. If, instead, the set of possible systems is infinite (e.g., when a system is identified in terms of the settings of continuous parameters), online learning can allow us to find the best such system efficiently.

Within this paper, we define an online learning system as a system that changes its behavior through interaction with its environment. A natural fit for this task are problem formulations from reinforcement learning [71]. Reinforcement learning is a branch of machine learning where agents (e.g., an IR system) interact with an environment (e.g., users) and learn by trying out actions (e.g., documents, news items, etc.) and observing rewards (e.g., interpret user actions as absolute or relative feedback). The full reinforcement learning problem also specifies states in which the environment can be in, and transitions between states, which may depend on the agent's action. In this paper, we focus on a subset of problems called bandit problems, where system actions do not affect future states. Initial work on taking state transitions into account has been conducted in the context of exploratory search [33] and session search [52].

Bandit problems are a natural fit for many online learning tasks in information retrieval, where characteristics of incoming users are independent of other users. One mapping to web search is shown in Fig. 2 (analogous mappings hold for other IR tasks, such as news recommendation, ad placement, etc.). Here, the system learns from user interactions, by taking actions (selecting documents or document rankings), and observing user feedback. Interactions are modeled in rounds or discrete timesteps, where in each timestep the agent may observe some context, generates an action, and observes and applies feedback. A crucial difference to learning in a supervised setting is that only feedback for selected actions is observed. The task of the learner is to optimize online performance, i.e., performance while learning. These two characteristics result in the exploration-exploitation challenge, because actions with unknown performance have to be explored to learn better solutions. An important benefit of reducing IR problems

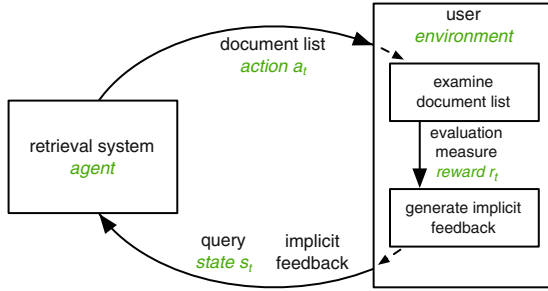


Fig. 2. Example formulation of search as a contextual bandit problem, with information retrieval terminology shown in black, and reinforcement learning terminology shown in green (Color figure online).

to bandit approaches is that the rich body of work on bandit approaches can be leveraged. At the same time, IR poses some unique challenges that further drive development in bandit research, such as approaches that work with relative feedback (discussed in Sect. 7).

Many types of bandit approaches have been developed. Here, we divide these approaches in terms of how they interpret feedback for learning. In this section, we focus on approaches with absolute feedback. We outline work in three areas, and show how the developed approaches relate to IR applications. We start from the non-contextual K -armed bandit setting, where the payoffs of available actions are independent. This is extended to the contextual setting, where context provides additional information on when an action may have high reward. Finally, we consider extensions to large or infinite action spaces. A detailed survey of bandit approaches and their analysis can be found in [9].

In the classic *K-armed bandit* setting, the learner has to select from a finite set of available actions. A simple approach that often works surprisingly well in practice is ϵ -greedy [77]. At each timestep, it explores with probability ϵ by randomly selecting an available action, and exploits the empirically best action with probability $1 - \epsilon$. Convergence guarantees are known for appropriate choice of ϵ . Another popular type of approach is UCB (upper confidence bound) [3]. It maintains estimates of the expected payoff for each available action, constructs confidence intervals around these estimates, and at each timestep selects the action with the highest upper confidence bound. Convergence guarantees exist for the stochastic setting, where payoff for each action is assumed to be independently sampled from a stationary distribution. An approach that does not rely on stochastic feedback is EXP [4]. Approaches of this type maintain a distribution over actions' expected payoff and sample from this distribution. Because of its stochastic nature, EXP has performance guarantees even in adversarial settings, where payoffs are selected by an adversary that competes with the learner. Recently, approaches based on Thompson Sampling have been shown to achieve good empirical performance [13]. This finding triggered much theoretical work to better understand properties of this approach analytically [1, 36, 62].

The approach works by maintaining a distribution over expected payoffs for all arms, and at each timestep sampling from this distribution and acting optimally according to the drawn sample.

A pioneering approach that applied bandit approaches to IR was proposed by [60]. The authors formulated the task of learning diverse rankings with bandit feedback. Assuming a user population with diverse information needs, the task is to learn rankings that satisfy as many users as possible (i.e., show at least one relevant document per intent). This problem was later generalized to the submodular bandit problem, where a set of items has to be selected to optimize submodular utility functions [46, 70, 78].

An extension of the classic K-armed bandit problem that is particularly relevant to IR problems is the *contextual bandit problem* (also known as bandits with side-information, associative bandits, and bandits with expert advice) [49]. Here, the learner is given additional information in each round, that can help identify the action to select. In an IR setting, this context information can consist of, for example, a user profile or history, a query, a website on which an ad must be placed, etc. Naively, K-armed bandit approaches can be applied to this setting by learning a separate bandit for each context. However, this approach results in a large increase in the amount of required exploration (all actions have to be explored sufficiently often in each context), and consequently a reduction in online performance. However, extensions to K-armed bandit approaches have been developed that efficiently generalize over contexts. For example, EXP4 generalizes over actions by transforming the exploration problem to exploration in some contextual policy space [4]. Langford and Zhang [47] extend ideas from ϵ -greedy to continuous contexts. LinUCB extends UCB to the contextual bandit setting by generalizing it to a linear reward model [49], and similar approaches are explored for Thompson Sampling [62]. A linear approach with submodular utility functions is proposed in [78].

Much of the work on contextual bandit approaches was informed by, and empirically validated on, IR problems such as news recommendation [49, 78], ad placement [48], vertical selection [20, 32], comment recommendation [53], ad format selection [73], and, most recently, spell correction in search queries [51]. These problems can be accurately modeled as contextual bandit problems with small action sets and high-dimensional context information, with absolute reward metrics such as clickthrough rate.

An orthogonal extension of bandit approaches is to consider large or infinite action spaces. In settings, where the number of actions is large, as is the case when searching large document collections, exploring all available actions is prohibitive. Approaches that tackle this problem exploit information about the similarity of actions. This information can be provided explicitly, often in the form of a tree structure [42]. Extensions of this work generalize to cases where properties of the underlying space are unknown and feedback stochastic [74]. Slivkins et al. [68] extend this approach to the ranked bandit setting, to learn diverse subsets of large or infinite action spaces.

In this section we discussed online learning approaches for IR. We focused on bandit approaches for learning in the K-armed and contextual setting, and briefly outlined approaches for learning in settings with large or infinite action spaces. The approaches described so far learn effectively in settings where reliable absolute feedback, such as clickthrough, can be observed. In the next section, we discuss bandit approaches that learn from relative feedback.

7 Online Learning from Relative Feedback

In many interactive IR systems, absolute reward may not accurately reflect user satisfaction or other target quantities, because they are too context dependent and noisy (cf., Sect. 4). For these settings, relative feedback methods have been developed, and have been shown to be substantially more robust (cf., Sect. 5). Naturally, we would like to use these relative metrics as feedback for online learning. While classic approaches (such as the bandit approaches discussed in Sect. 6) focused on absolute feedback settings, the first relative approaches have been proposed recently. We give a brief overview of these in this section. A thorough review of relative bandit algorithms (also called preference-based multiarmed bandits) was recently published by Busa-Fekete and Hüllermeier [10].

Supervised learning approaches for learning from relative feedback go back to at least [34]. In IR, this approach has been very successfully applied to supervised learning to rank problems, where expert relevance labels can be interpreted as relative feedback. However, supervised approaches do not address the exploration-exploitation challenge, and directly applying supervised approaches to interaction data is very susceptible to noise and bias. If applied to learn in a batch setting from exploration data, very high levels of exploration would be required to combat bias [25]. Dueling bandit approaches naturally address the exploration-exploitation challenge, while working with relative feedback.

The K-armed dueling bandit problem was first formulated by Yue et al. [81, 83], and was directly motivated by the need to learn from relative feedback in IR settings. It generalizes multiarmed bandit problems to settings where absolute performance cannot be quantified, but comparisons between two arms can be made. These comparisons can be stochastic, such that the a better arm i has a probability of winning a comparison against a worse arm j of $p_{ij} > 0.5$. They propose an approach to solving this problem, called Interleaved Filter (IF), which works in rounds in which it eliminates an arm when it is proven to have low performance. Since then, new dueling bandit approaches have been developed that substantially improve over both the empirical performance of IF, and over its theoretical guarantees. For example, Beat-the-Mean compares each arm to the sampled mean of all arms [80].

The main challenge addressed by dueling bandit approaches is to select the arms to compete in each round such that the competition quickly focuses on the best arms, to avoid excessive exploration of bad arms. Zoghi et al. [85] employs a relative UCB-style approach, such that it always selects the arm with the highest confidence bound as one competitor, and plays it against the arm that has the

best chance of beating it. A strategy based on Thompson Sampling is proposed in [84]. Another very recent approach is by Ailon et al. [2], who provide several reductions from dueling bandits to classic cardinal bandits.

In addition to the K-armed dueling bandit problem, Yue and Joachims [79] proposed a contextual problem formulation, resulting in a generic dueling bandit formulation. In this formulation, the learner has to optimize a linear function in d dimensions, using only relative feedback about the relative performance of two such solutions. A stochastic gradient descent approach to solving this problem is the Dueling Bandits Gradient Descent (DBGD) [79]. Briefly, it learns by interacting with the environment in rounds, and observing relative feedback as follows. At all times, the learner maintains a “current best” solution w_t (a solution is a weight vector for linear weighted combination of context features). In each round, it generates a “challenger” w'_t , by randomly sampling from a unit sphere around w_t . Then, w_t and w'_t are compared (e.g., when learning rankings, this could be done using interleaving). If w'_t wins the comparison, w_t is updated by a learning step in the direction of w'_t . If the solution space is convex, this approach is guaranteed to converge [79].

DBGD can be directly applied to e.g., online learning to rank settings, and was empirically validated on such a task [79]. However, it is more generally applicable, as it makes no specific assumptions about how solutions are compared, as long as assumptions of the algorithm regarding their stochastic characteristics hold. Hofmann et al. [29] demonstrated that, by taking structure into account, substantially better online and offline performance can be achieved in specific applications. For the task of online learning to rank from interleaved comparisons, an approach called Candidate Pre-Selection (CPS) was proposed. It leverages the exploration that is a side-effect of probabilistic interleave (see Sect. 5), to evaluate new ranker candidates. In comparison to DBGD, which explores uniformly around the current best solution, CPS uses offline estimates derived from exploration data to focus on the most promising candidates. The resulting approach learns significantly faster than the structure oblivious approach, and is particularly robust to noisy feedback [29]. Schuth et al. [66] that dueling bandit approaches can be successfully applied to learning the parameters of non-linear ranking functions.

In this and the previous section, we have provided a summary of the many online learning approaches that can enable interactive retrieval systems to learn directly from interactions with their users. Interestingly, the unique challenges posed by IR applications have motivated many recent advances in e.g., contextual and dueling bandit approaches. As these make their way into more and more interactive IR systems, we can expect to discover and solve new challenges.

8 Conclusion

In this paper we have presented an overview of techniques for online experimentation for IR. With the increase in web-scale IR systems, controlled experiments have been adapted to deal with the challenges of scale and complexity that these

systems present. As well as moving insights into experimentation methodology into practical settings, new methods for measurement and learning have been developed that can in turn benefit IR research.

The basis of online experimentation for IR naturally are controlled experiments. In Sect. 2 we motivated why the causal relations they let us infer are crucial for systems that learn how to act, or interact, with their users. After motivating the need for controlled experimentation, we introduced the most well-known technique, A/B testing. A/B testing is the technique that is the most general, but has limitations in terms of the scalability of comparisons. This gap is filled by methods for estimating online performance from exploration data, as discussed in Sect. 3. In Sect. 4 we briefly discussed recent trends in measuring online performance of IR systems. Section 5 concluded our discussion of online evaluation, by introducing interleaved comparison methods, which allow within-subject controlled experiments.

The first sections of this paper focused on online learning for IR. Online learning goes one step beyond the previously discussed online evaluation approaches where the comparisons to be performed were selected manually. Online learning using bandit approaches in particular can automatically select the required evaluations or comparisons in order to optimize online performance. A key challenge addressed by bandit approaches is the trade-off between exploring new solutions to obtain accurate performance estimates, and exploiting solutions with known high performance. The resulting approaches are especially valuable for online learning in IR systems, as they achieve high online performance while learning.

Following on from this overview article, many of the presented approaches can be tried out in the experimental framework *lerot* [65]. Using simulations of online interactions, online evaluation and learning approaches can be compared and developed further.

Online evaluation and learning have only recently been introduced to the IR community, and form a growing area of research within this community. Many open questions remain to be addressed. From the perspective of deploying online evaluation and learning approaches, we need to better understand the impact of exploration on the user experience. While exploration allows learning, and therefore improves system performance in the long run, it is not yet well-understood how users are affected in the short run, and how potential risks can be mitigated. Particularly valuable are exploration schemes that limit the risk for individual users. On the other hand, we need to better understand how to effectively explore in applications with large action spaces. The more we know about the solution space of a given IR problem, the more effectively we can design exploration schemes that use this structure to quickly focus on the most promising areas of the solution space.

Online experimentation has been embraced by owners of large web properties, and is a key part of the development process in these companies. In the research community, online experimentation seems to see somewhat slower adoption. Is one reason the difficulty in obtaining data or running experiments in an online setting? Exploration data may be a key to enabling wider participation in online

experimentation. Another promising initiative is the CLEF living labs initiative, which brings together IR researchers and search engine operators, by providing a shared platform for online experimentation [6].

What questions can we study using online experimentation for IR? The methods presented in this article build on and complement the traditional toolset of IR experimentation. Online experimentation can expand IR research from small-scale and short-term lab studies to a wide range of naturalistic experiments. This will allow us to gain new insights into information seeking behavior, and into how retrieval systems can best address these.

Online learning for IR can transform the way in which IR systems are currently developed. By learning directly from user interactions, they can quickly adapt to changing user behavior and expectations. We will move away from developing systems for which behavior is completely specified before deployment, and will move towards defining a space of possible solutions. Online evaluation and online learning to rank will drive this development, towards IR systems that learn directly from their users.

References

1. Agrawal, S., Goyal, N.: Analysis of thompson sampling for the multi-armed bandit problem. In: COLT 2012 (2012)
2. Ailon, N., Karnin, Z., Joachims, T.: Reducing dueling bandits to cardinal bandits. In: ICML 2014 (2014)
3. Auer, P., Cesa-Bianchi, N., Fischer, P.: Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.* **47**(2–3), 235–256 (2002)
4. Auer, P., Cesa-Bianchi, N., Freund, Y., Schapire, R.E.: The nonstochastic multi-armed bandit problem. *SIAM J. Comput.* **32**(1), 48–77 (2002)
5. Babbie, E.R.: *The Practice of Social Research*, 13th edn. Cengage Learning, Boston (2012)
6. Balog, K., Kelly, L., Schuth, A.: Head first: Living labs for ad-hoc search evaluation. In: CIKM 2014 (2014)
7. Bendersky, M., Garcia-Pueyo, L., Harmsen, J., Josifovski, V., Lepikhin, D.: Up next: Retrieval methods for large scale related video suggestion. In: KDD 2014 (2014)
8. Bottou, L., Chickering, J., Portugaly, E., Ray, D., Simard, P., Snelson, E.: Counterfactual reasoning and learning systems: The example of computational advertising. *J. Mach. Learn. Res.* **14**(1), 3207–3260 (2013)
9. Bubeck, S., Cesa-Bianchi, N.: Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Found. Trends Mach. Learn.* **5**(1), 1–122 (2012)
10. Busa-Fekete, R., Hüllermeier, E.: A survey of preference-based online learning with bandit algorithms. In: Auer, P., Clark, A., Zeugmann, T., Zilles, S. (eds.) ALT 2014. LNCS, vol. 8776, pp. 18–39. Springer, Heidelberg (2014)
11. Carterette, B.: Statistical significance testing in information retrieval: Theory and practice. In: ICTIR 2013 (2013)
12. Chakraborty, S., Radlinski, F., Shokouhi, M., Baecke, P.: On correlation of absence time and search effectiveness. In: SIGIR 2014, pp. 1163–1166 (2014)
13. Chapelle, O., Li, L.: An empirical evaluation of thompson sampling. In: NIPS 2011, pp. 2249–2257 (2011)

14. Chapelle, O., Zhang, Y.: A dynamic bayesian network click model for web search ranking. In: WWW 2009, pp. 1–10 (2009)
15. Chapelle, O., Joachims, T., Radlinski, F., Yue, Y.: Large-scale validation and analysis of interleaved search evaluation. *ACM Trans. Inf. Syst.* **30**(1), 6:1–6:41 (2012)
16. Chuklin, A., Schuth, A., Hofmann, K., Serdyukov, P., de Rijke, M.: Evaluating aggregated search using interleaving. In: CIKM 2013 (2013)
17. Chuklin, A., Schuth, A., Zhou, K., de Rijke, M.: A comparative analysis of interleaving methods for aggregated search. *ACM Trans. Inf. Syst.* (2014)
18. Craswell, N., Zoeter, O., Taylor, M., Ramsey, B.: An experimental comparison of click position-bias models. In: WSDM 2008, pp. 87–94 (2008)
19. Deng, A., Xu, Y., Kohavi, R., Walker, T.: Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. In: WSDM 2013, pp. 123–132 (2013)
20. Diaz, F.: Adaptation of offline vertical selection predictions in the presence of user feedback. In: SIGIR 2009, pp. 323–330 (2009)
21. Dupret, G., Lalmas, M.: Absence time and user engagement. In: WSDM 2013, p. 173. ACM Press, New York, February 2013
22. Granka, L.A., Joachims, T., Gay, G.: Eye-tracking analysis of user behavior in www search. In: SIGIR 2004, pp. 478–479 (2004)
23. Guan, Z., Cutrell, E.: An eye tracking study of the effect of target rank on web search. In: CHI 2007, pp. 417–420 (2007)
24. Hassan, A., White, R.W.: Personalized models of search satisfaction. In: CIKM 2013, pp. 2009–2018 (2013)
25. Hofmann, K., Whiteson, S., de Rijke, M.: Balancing exploration and exploitation in learning to rank online. In: Clough, P., Foley, C., Gurrin, C., Jones, G.J.F., Kraaij, W., Lee, H., Mudooh, V. (eds.) *ECIR 2011*. LNCS, vol. 6611, pp. 251–263. Springer, Heidelberg (2011)
26. Hofmann, K., Whiteson, S., de Rijke, M.: A probabilistic method for inferring preferences from clicks. In: CIKM 2011, pp. 249–258 (2011)
27. Hofmann, K., Behr, F., Radlinski, F.: On caption bias in interleaving experiments. In: CIKM 2012, pp. 115–124. ACM Press (2012)
28. Hofmann, K., Whiteson, S., de Rijke, M.: Estimating interleaved comparison outcomes from historical click data. In: CIKM 2012, pp. 1779–1783 (2012)
29. Hofmann, K., Whiteson, S., de Rijke, M.: Balancing exploration and exploitation in listwise and pairwise online learning to rank for information retrieval. *Inf. Retrieval J.* **16**(1), 63–90 (2013)
30. Hofmann, K., Whiteson, S., de Rijke, M.: Fidelity, soundness, and efficiency of interleaved comparison methods. *ACM Trans. Inf. Syst.* **31**(4), 1–43 (2013)
31. Hofmann, K., Mitra, B., Radlinski, F., Shokouhi, M.: An eye-tracking study of user interactions with query auto completion. In: CIKM 2014 (2014)
32. Jie, L., Lamkhede, S., Sapra, R., Hsu, E., Song, H., Chang, Y.: A unified search federation system based on online user feedback. In: KDD 2013, pp. 1195–1203 (2013)
33. Jin, X., Sloan, M., Wang, J.: Interactive exploratory search for multi page search results. In: WWW 2013, pp. 655–666 (2013)
34. Joachims, T.: Optimizing search engines using clickthrough data. In: KDD 2002, pp. 133–142 (2002)
35. Joachims, T., Granka, L., Pan, B., Hembrooke, H., Radlinski, F., Gay, G.: Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *ACM Trans. Inf. Syst.* **25**(2), 1–26 (2007)

36. Kaufmann, E., Korda, N., Munos, R.: Thompson sampling: an asymptotically optimal finite-time analysis. In: Bshouty, N.H., Stoltz, G., Vayatis, N., Zeugmann, T. (eds.) ALT 2012. LNCS, vol. 7568, pp. 199–213. Springer, Heidelberg (2012)
37. Kazai, G., Kamps, J., Koolen, M., Milic-Frayling, N.: Crowdsourcing for book search evaluation: Impact of hit design on comparative system ranking. In: SIGIR 2011, pp. 205–214 (2011)
38. Kelly, D.: Methods for evaluating interactive information retrieval systems with users. *Found. Trends Inf. Retrieval* **3**(1–2), 1–224 (2009)
39. Kelly, D., Teevan, J.: Implicit feedback for inferring user preference: a bibliography. *SIGIR Forum* **37**(2), 18–28 (2003)
40. Kelly, D., Gyllstrom, K., Bailey, E.W.: A comparison of query and term suggestion features for interactive searching. In: SIGIR 2009, p. 371. ACM Press, New York, July 2009
41. Kim, Y., Hassan, A., White, R.W., Zitouni, I.: Modeling dwell time to predict click-level satisfaction. In: WSDM 2014, pp. 193–202. ACM, New York (2014)
42. Kleinberg, R., Slivkins, A., Upfal, E.: Multi-armed bandits in metric spaces. In: STOC 2008. ACM Press (2008)
43. Kohavi, R., Longbotham, R., Sommerfield, D., Henne, R.M.: Controlled experiments on the web: survey and practical guide. *Data Min. Knowl. Disc.* **18**(1), 140–181 (2009)
44. Kohavi, R., Deng, A., Frasca, B., Longbotham, R., Walker, T., Xu, Y.: Trustworthy online controlled experiments: Five puzzling outcomes explained. In: KDD 2012, pp. 786–794. ACM, New York (2012)
45. Kohavi, R., Deng, A., Frasca, B., Walker, T., Xu, Y., Pohlmann, N.: Online controlled experiments at large scale. In: KDD 2013, pp. 1168–1176. ACM, New York (2013)
46. Kohli, P., Salek, M., Stoddard, G.: A fast bandit algorithm for recommendation to users with heterogeneous tastes. In: AAAI 2013 (2013)
47. Langford, J., Zhang, T.: The epoch-greedy algorithm for multi-armed bandits with side information. In: NIPS 2008, pp. 817–824 (2008)
48. Langford, J., Strehl, A., Wortman, J.: Exploration scavenging. In: ICML 2008, pp. 528–535 (2008)
49. Li, L., Chu, W., Langford, J., Schapire, R.E.: A contextual-bandit approach to personalized news article recommendation. In: WWW 2010, pp. 661–670 (2010)
50. Li, L., Chu, W., Langford, J., Wang, X.: Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In: WSDM 2011, pp. 297–306 (2011)
51. Li, L., Chen, S., Kleban, J., Gupta, A.: Counterfactual estimation and optimization of click metrics for search engines (2014). arXiv preprint [arXiv:1403.1891](https://arxiv.org/abs/1403.1891)
52. Luo, J., Zhang, S., Yang, H.: Win-win search: Dual-agent stochastic game in session search. In: SIGIR 2014, pp. 587–596. ACM (2014)
53. Mahajan, D.K., Rastogi, R., Tiwari, C., Mitra, A.: LogUCB: An explore-exploit algorithm for comments recommendation. In: CIKM 2012, pp. 6–15 (2012)
54. Pearl, J.: *Causality: Models, Reasoning and Inference*, vol. 29. Cambridge University Press, Cambridge (2000)
55. Pearl, J.: An introduction to causal inference. *Int. J. Biostatistics* **6**(2) (2010)
56. Precup, D., Sutton, R.S., Singh, S.P.: Eligibility traces for off-policy policy evaluation. In: ICML 2000, pp. 759–766 (2000)
57. Radlinski, F., Craswell, N.: Comparing the sensitivity of information retrieval metrics. In: SIGIR 2010, pp. 667–674 (2010)

58. Radlinski, F., Craswell, N.: Optimized interleaving for online retrieval evaluation. In: WSDM 2013 (2013)
59. Radlinski, F., Joachims, T.: Minimally invasive randomization for collecting unbiased preferences from clickthrough logs. In: AAAI 2006, p. 1406 (2006)
60. Radlinski, F., Kleinberg, R., Joachims, T.: Learning diverse rankings with multi-armed bandits. In: ICML 2008, pp. 784–791. ACM (2008)
61. Radlinski, F., Kurup, M., Joachims, T.: How does clickthrough data reflect retrieval quality?. In: CIKM 2008, pp. 43–52 (2008)
62. Russo, D., Roy, B.V.: An information-theoretic analysis of thompson sampling. CoRR, abs/1403.5341 (2014). URL <http://arxiv.org/abs/1403.5341>
63. Sanderson, M.: Test collection based evaluation of information retrieval systems. Found. Trends Inf. Retrieval 4(4), 247–375 (2010)
64. Scholer, F., Shokouhi, M., Billerbeck, B., Turpin, A.: Using clicks as implicit judgments: expectations versus observations. In: Macdonald, C., Ounis, I., Plachouras, V., Ruthven, I., White, R.W. (eds.) ECIR 2008. LNCS, vol. 4956, pp. 28–39. Springer, Heidelberg (2008)
65. Schuth, A., Hofmann, K., Whiteson, S., de Rijke, M.: Lerot: an online learning to rank framework. In: LivingLab 2013, pP. 23–26. ACM (2013)
66. Schuth, A., Sietsma, F., Whiteson, S., de Rijke, M.: Optimizing base rankers using clicks. In: de Rijke, M., Kenter, T., de Vries, A.P., Zhai, C.X., de Jong, F., Radinsky, K., Hofmann, K. (eds.) ECIR 2014. LNCS, vol. 8416, pp. 75–87. Springer, Heidelberg (2014)
67. Schuth, A., Sietsma, F., Whiteson, S., Lefortier, D., de Rijke, M.: Multileaved comparisons for fast online evaluation. In: CIKM 2014 (2014)
68. Slivkins, A., Radlinski, F., Gollapudi, S.: Ranked bandits in metric spaces: learning diverse rankings over large document collections. J. Mach. Learn. Res. 14(1), 399–436 (2013)
69. Song, Y., Shi, X., Fu, X.: Evaluating and predicting user engagement change with degraded search relevance. In: WWW 2013, pp. 1213–1224 (2013)
70. Streeter, M., Golovin, D., Krause, A.: Online learning of assignments. In: NIPS 2009, pp. 1794–1802 (2009)
71. Sutton, R.S., Barto, A.G.: Introduction to Reinforcement Learning. MIT Press, Cambridge (1998)
72. Tang, D., Agarwal, A., O’Brien, D., Meyer, M.: Overlapping experiment infrastructure: More, better, faster experimentation. In: KDD 2010, pp. 17–26 (2010)
73. Tang, L., Rosales, R., Singh, A., Agarwal, D.: Automatic ad format selection via contextual bandits. In: CIKM 2013, pp. 1587–1594 (2013)
74. Valko, M., Carpentier, A., Munos, R.: Stochastic simultaneous optimistic optimization. In: ICML 2013, pp. 19–27 (2013)
75. Voorhees, E.M., Harman, D.K.: TREC: Experiment and Evaluation in Information Retrieval. Digital Libraries and Electronic Publishing. MIT Press, Cambridge (2005)
76. Wang, K., Walker, T., Zheng, Z.: PSkip: estimating relevance ranking quality from web search clickthrough data. In: KDD 2009, pp. 1355–1364 (2009)
77. Watkins, C.J.C.H.: Learning from delayed rewards. Ph.D. thesis, University of Cambridge (1989)
78. Yue, Y., Guestrin, C.: Linear submodular bandits and their application to diversified retrieval. In: Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., Weinberger, K. (eds.) NIPS 2011, pp. 2483–2491 (2011)
79. Yue, Y., Joachims, T.: Interactively optimizing information retrieval systems as a dueling bandits problem. In: ICML 2009, pp. 1201–1208 (2009)

80. Yue, Y., Joachims, T.: Beat the mean bandit. In: ICML 2011 (2011)
81. Yue, Y., Broder, J., Kleinberg, R., Joachims, T.: The K-armed dueling bandits problem. In: COLT 2009 (2009)
82. Yue, Y., Patel, R., Roehrig, H.: Beyond position bias: examining result attractiveness as a source of presentation bias in clickthrough data. In: WWW 2010, pp. 1011–1018 (2010)
83. Yue, Y., Broder, J., Kleinberg, R., Joachims, T.: The K-armed dueling bandits problem. *J. Comput. Syst. Sci.* **78**(5), 1538–1556 (2012)
84. Zoghi, M., Whiteson, S.A., de Rijke, M., Munos, R.: Relative confidence sampling for efficient on-line ranker evaluation. In: WSDM 2014, pp. 73–82 (2014)
85. Zoghi, M., Whiteson, S.A., Munos, R., de Rijke, M.: Relative upper confidence bound for the K-armed dueling bandit problem. In: ICML 2014 (2014)

Information Retrieval

8th Russian Summer School, RuSSIR 2014, Nizhniy

Novgorod, Russia, August 18-22, 2014, Revised

Selected Papers

Braslavski, P.; Karpov, N.; Worring, M.; Volkovich, Y.;

Ignatov, D.I. (Eds.)

2015, XIV, 359 p. 125 illus., Softcover

ISBN: 978-3-319-25484-5