Competitive Search

Oren Kurland
kurland@technion.ac.il
Technion — Israel Institute of Technology
Haifa, Israel

Moshe Tennenholtz
moshet@ie.technion.ac.il
Technion — Israel Institute of Technology
Haifa, Israel

ABSTRACT

The Web is a canonical example of a *competitive search setting* that includes document authors with ranking incentives: their goal is to promote their documents in rankings induced for queries. The incentives affect some of the corpus dynamics as the authors respond to rankings by applying strategic document manipulations. This well known reality has deep consequences that go well beyond the need to fight spam. As a case in point, researchers showed using game theoretic analysis that the probability ranking principle is not optimal in competitive retrieval settings; specifically, it leads to reduced topical diversity in the corpus. We provide a broad perspective on recent work on competitive retrieval settings, argue that this work is the tip of the iceberg, and pose a suite of novel research directions; for example, a general game theoretic framework for competitive search, methods of learning-to-rank that account for post-ranking effects, approaches to automatic document manipulation, addressing societal aspects and evaluation.

CCS CONCEPTS

• **Information systems → Information retrieval**; Search engine architectures and scalability; **Adversarial retrieval**;

KEYWORDS

competitive search; game theory; search engine optimization

ACM Reference Format:
Oren Kurland and Moshe Tennenholtz. 2022. Competitive Search. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22), July 11–15, 2022, Madrid, Spain.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3477495.3532771

1 INTRODUCTION

Ad hoc retrieval is a fundamental and long studied task in the information retrieval field: retrieving documents relevant to an information need expressed by a query. The vast majority of work on devising retrieval methods assumes a static setting; that is, a (single) query and a corpus snapshot. The prominent example where the retrieval setting is dynamic is interactive retrieval where the user interacts with the search system, e.g., by providing relevance feedback [55, 110, 115]. Still, the corpus is assumed to be fixed.

While it is a fact that retrieval is performed against a corpus snapshot (via an index), the corpus can change throughout time. Furthermore, some of the corpus changes can be driven by rankings induced for queries: if document authors, henceforth referred to as publishers, are not satisfied with the positions of their documents in ranked lists, they might modify them so as to potentially improve their future ranking. Therefore, these publishers are ranking incentivized and their incentives drive corpus dynamics. The search setting then becomes *competitive*, as publishers compete for high rankings.

Document manipulations intended to improve future rankings are often referred to as search engine optimization (SEO) [53]. SEO is pervasive over the Web where some publishers compete for high rankings; e.g., for queries with a commercial intent. One reason is monetization: users pay much attention to top retrieved results [64], and engagement with the results in some cases translates to financial rewards. Hence, the Web (and more specifically, organic search) is a canonical example of a competitive retrieval setting. Additional examples of competitive retrieval settings with SEO phenomena are product search in e-commerce platforms [47, 117] and sponsored search [6, 14].

The vast majority of past work on adversarial search in competitive retrieval settings focused on spam [27]. Spam is perhaps the main example of black hat SEO: actions that are not legitimate as they hurt the search echosystem. There is a large body of work on addressing different types of spam (e.g., content or link-based) [1, 17, 26, 28–30, 39, 54, 60, 65, 83, 111]. Furthermore, there is work on penalizing low quality documents in rankings [13].

White hat SEO [53], on the other hand, is a suite of legitimate document modification strategies intended to promote documents in rankings. While legitimate, the modifications, often performed in response to induced rankings, affect the corpus. As it turns out, these modifications can hurt the search echosystem in the long run as was recently shown [10, 11]. Specifically, the theoretical grounds of ranking in standard non-competitive retrieval settings, namely the probability ranking principle (PRP) [88], was shown to be sub-optimal in competitive settings [10, 11]. The PRP, due to white hat SEO, leads to reduced topical diversity in the corpus [10, 11]. In other words, the ability to satisfy future information needs is hurt due to publishers' responses to rankings induced based on the PRP. It is not only the PRP that breaks down in competitive settings: some axioms which were proposed for retrieval in standard settings are not suitable for competitive retrieval settings; e.g., increased frequency of query terms in a document could be due to strategic manipulations, and hence, should not necessarily translate to increased retrieval scores as is the case in non-competitive settings [38].

A recent line of work [10, 11, 49, 49, 50, 87, 104] addresses the competitive retrieval setting, specifically in terms of white hat SEO.

SIGIR '22, July 11–15, 2022, Madrid, Spain

ACM ISBN 978-1-4503-8732-3/22/07…$15.00
https://doi.org/10.1145/3477495.3532771

Novel research questions are raised, novel approaches are devised — specifically, based on game theory — and evaluation is performed using organized ranking competitions. However, this seems to be only the tip of the iceberg.

In this paper we provide a broad perspective on different aspects and issues in competitive retrieval settings. We start by analyzing a recently proposed initial game theoretical modeling of the competitive setting [10, 11]. Game theory is well suited to modeling dynamic settings with incentivized players. In competitive retrieval settings, the publishers are the players. We argue for the need for a much more rigorous game-theory-based framework that will allow to contrast retrieval methods and principles. Furthermore, such a framework will have to address various aspects of ad hoc retrieval: ranking based on multiple criteria and not only relevance (e.g., search results diversification [96]) and interactive retrieval [110, 115].

We next turn to examine ranking functions in competitive retrieval settings and the way they are learned (i.e., learning-to-rank). Given the findings about the sub-optimality of the PRP [10, 11], and the need to account for long term corpus effects of rankings, using the standard approach of a loss function targeting relevance effectiveness falls short; that is, there is no account for post-ranking effects. We discuss the challenge of optimizing for both short term relevance effectiveness and long term corpus effects. We then discuss the need to make current relevance estimates, classical (e.g., Okapi BM25 [89]) and modern (e.g., [52, 72, 77]), robust to ranking-incentivized document manipulations.

As a next step we focus on the publisher perspective. In an era with dramatic improvements in language generation abilities based on pre-trained language models (e.g., BERT [33], XLNet [112], GPT3 [20]), automatic content generation is becoming quite pervasive. We examine the challenge of automatic document modifications for improved rankings (i.e., automatic white hat SEO). The importance of addressing this challenge is three fold. First, documents of publishers with no access to such technology will essentially be penalized in rankings, due to being not competitive; this will lead to fairness issues [25]. Second, evaluation as we discuss below is a hard challenge in competitive settings. Hence, the ability of automatically modifying documents and even creating them is important for extending evaluation opportunities. Third, there is a large body of work in several research communities other than IR on adversarial attacks on neural methods which drove forward work on devising methods that are more robust to adversarial attacks [35, 57, 58, 62, 71, 84, 100, 103, 107]. The crypto community, for example, has seen throughout the years many publications of attacks on encryption algorithms and these led to improved encryption methods. This state-of-affairs is not the case in the information retrieval community, although potentially, the largest example of an adversarial and competitive setting is the Web.

Additional important aspect that we discuss is societal effects in the competitive retrieval setting. Recent work demonstrated the potential ability of publishers to use the search engine as a platform to promote corpus changes due to a herding phenomenon [50]. That is, it was shown both theoretically [87] and empirically [50, 87] that publishers tend to mimic content in documents previously highly ranked for queries of interest. The rationale is that induced rankings are the only signal about the undisclosed ranking function. Hence, highly ranked documents that manifest undesirable phenomenon (e.g., fake news) can badly influence the corpus. We discuss the need to develop methods to identify such potential situations.

We then move to discuss evaluation in competitive retrieval settings. To evaluate the effectiveness of a novel ranking function, one should consider in the competitive retrieval setting the responses of publishers' to rankings induced by the function, namely, document modifications. Hence, the standard evaluation approach of using a static corpus snapshot does not allow to perform proper evaluation in dynamic retrieval settings.

The issue of evaluation is in our opinion a major challenge in addressing competitive retrieval settings. It is perhaps one of the main reasons for why these settings were not addressed beyond work on detecting and addressing spam. First, query logs of commercial search engines operating over the Web, which is the canonical example of a competitive retrieval setting, are proprietary. Second, even with access to such logs, isolating specific responses to induced rankings, and more generally, focusing on specific phenomena is an extremely hard challenge; there are various factors that drive the dynamics of the Web which are not necessarily ranking incentives. In recent work on competitive retrieval [49, 50, 87], small scale controlled ranking competitions held between students served for evaluation. We discuss this type of competitions and describe how, in our opinion, evaluation should be extended and generalized.

We describe a suite of macro-level research directions (RDs) throughout this work. Each of them is essentially a strategic research avenue to explore which by itself entails many other research directions. We took care to balance the granularity of discussion of the RDs where a major motivation was to provide as broad perspective as possible.

To summarize the importance of addressing competitive retrieval, we note the following. As mentioned above, the theoretical underpinning of ad hoc retrieval methods (the PRP [88]) "breaks down" in competitive settings due to lack of accounting for post-retrieval effects. Other theoretical frameworks (e.g., the axiomatic approach [38]) will also have to be (considerably) changed to address ranking incentives. Existing relevance estimates are not robust to strategic manipulation whether these are "classic" (e.g., Okapi BM25 [89]) or "modern" (i.e., neural [52, 72, 77]). A case in point, simple keyword stuffing of query terms in a document can bias relevance estimates in unwarranted ways. At a more macro level, learning-to-rank for both short term relevance effectiveness and long-term corpus effects is a highly novel and intriguing research agenda to pursue. Various aspects and flavors of ad hoc retrieval have to be reconsidered for competitive settings including results diversification and interactive retrieval. Addressing competitive retrieval calls for a completely new suite of techniques based on game theory. There are societal aspects involved due to the ability of publishers to affect content trends in the corpus. And, evaluation in the face of corpus dynamics and responses of publishers to rankings cannot be based on the standard approach of utilizing a static corpus snapshot.

2 SCOPE

Our focus in this paper is on ad hoc retrieval in competitive retrieval settings where publishers' responses, in the form of document manipulations, are "legitimate" — i.e., white hat search engine optimization (SEO). Spamming, and more generally black hat SEO [1, 27], which has been the focus of most prior work on adversarial retrieval, is out of the scope of this paper.

In sponsored search [6, 14, 36], both relevance and monetization considerations of the search engine determine rankings. Balancing these two criteria [36] and addressing adversarial effects, specifically, using game theoretic approaches, was the focus of several studies [6, 14]. In contrast, our focus is on *organic search* — a setting where monetization considerations are not applied by the search engine.

As already noted, competitive product search is an integral part of e-commerce platforms where the financial incentive is clear and direct [47, 117]. Similarly to sponsored search, ranking is based on relevance and monetization. Hence, this is also outside the scope of this paper, despite being a highly interesting and challenging task. There is also work on the dynamic search settings in two sided markets [40, 41], but the focus is not on content and relevance estimation as is the case in ad hoc retrieval.

The competitive landscape of recommendation systems has also been addressed lately using game theory [12, 80]. More generally, the interests and consequences thereof of multiple stakeholders in a recommendation setting — which is often viewed as a two sided market with a mediator — have been addressed in several reports [76]. There are close connections between the competitive search and recommendation settings: a user is provided with search results or recommendations by a mediator that ranks the content provided by incentivized content providers. Indeed, a unified perspective for the need to use game theory for competitive ad hoc retrieval, recommendation and other data science tasks has recently been presented [102]. Most aspects that we discuss, as well as almost all research directions/questions pertaining to ad hoc retrieval that we describe, were not discussed in this work. While we focus on ad hoc retrieval in this paper, some of the directions we suggest also apply to recommendation systems.

There is a line of work on using game theory to compare reward mechanisms for content creation by ad hoc contributors [45, 46]; e.g., crowd workers or users answering questions or commenting on Web sites. The content contributors can decide whether or not to contribute on a per-task basis which is major difference, among others, with the ad hoc retrieval setting we focus on here.

Our focus is on the competition between publishers and its effects on the corpus given a single search engine. There has also been work on competition between search engines which we do not address here [59].

3 OVERVIEW

The search engine is a *mediator*. Its ranking decisions can drive some of the corpus dynamics given publishers' ranking incentives. Indeed, some publishers can be viewed as strategic *players*. Game theory provides effective tools to modeling and reasoning about competitive settings driven by incentives. In Section 4 we briefly describe a recently proposed game theoretic approach to modeling the competitive ad hoc retrieval setting. We continue the section by portraiting a palette of open research questions and directions needed to further establish a rigorous game-theory-based framework for these settings.

In Section 5 we discuss a novel long-term corpus effects aspect that should be considered when devising ranking functions in competitive settings. Its integration with standard short term search effectiveness optimization is a fundamental challenge that we focus on.

We then move to consider the publisher perspective in Section 6. The dramatic advances in language-generation abilities based on pre-trained language models (e.g., BERT [33], XLNet [112], GPT3 [20], etc.), and the recent demonstration of the potential to automatically shape documents' content so as to effectively improve their rankings [49], give rise to a suite of fundamental questions.

The corpus dynamics due to ranking incentives can also have societal effects; e.g., the spread of misinformation or reduced topical diversity [50]. We discuss these societal aspects in Section 7.

The empirical study of ranking-incentives-based phenomena is extremely challenging. In Section 8 we survey some recent attempts to establish evaluation frameworks and describe important future directions to consider on this front.

4 A GAME THEORETIC FRAMEWORK

As described above, some of the competitive retrieval setting dynamics is driven by incentivized publishers who respond to induced rankings. Existing formal and theoretical frameworks of information retrieval cannot account for this dynamics: they are mainly based on searching a corpus snapshot without accounting for post-ranking corpus effects. Game theory, on the other hand, provides convenient tools for modeling and reasoning about such dynamics. For example, Ben Basat et al. [10, 11] modeled the competitive retrieval setting using game theory and showed that the probability ranking principle [88] is sub-optimal. Later, Raifer et al. [87] used game theory to explain the empirically observed strategic responses of ranking-incentivized publishers to induced rankings.

In what follows, we first introduce game theory background in the context of the ad hoc retrieval task (Section 4.1). Then, we shed light on what we view as much needed research directions for establishing a rigorous theoretical framework for retrieval in competitive retrieval settings (Section 4.2). Such framework will allow, for example, to contrast retrieval principles and methods in a methodological way.

4.1 Game Theory for Ad Hoc Retrieval

Ben Basat et al. [10, 11] defined ad hoc (query based) ranking games as follows. Players are publishers who write documents. A document can discuss a single topic or multiple topics. The topic(s) a publisher selects for a document is her action as a player in the game. A query is about a single topic and the distribution of incoming queries over topics is known. The relevance of a document to a query is also assumed to be known — i.e., this is an oracle setting, a point we re-visit below. A publisher is rewarded if and only if her document is the highest ranked for a query. This assumption corresponds to the fact that most user attention when browsing search results is paid to the highest ranked documents [64]. The utility (reward) of the publisher is 1 if the document is single topic and discusses the query topic. If the document is multi-topic, the utility is in $[0, 1]$ if one of the topics it discusses is the query topic[1]. Publishers respond to induced rankings by potentially changing the topic (proportions) for their documents.

A basic question that emerges is the state-of-affairs of the game in the long run; specifically, whether the game converges in terms of the documents the publishers produce. A Nash *equilibrium* in a game is a state where no player has an incentive to deviate from the actions she applies. Actions can be atomic (a.k.a., *pure*) or *mixed* — distributions over atomic actions. According to the celebrated result of Nash [82], a game with a finite number of players where each has a finite number of actions she can apply has a mixed-strategy equilibrium. A strategy is the action (or distribution thereof) employed by a player.

Games can potentially reach different equilibria wherein players have different (expected[2]) utilities. One way to quantify equilibria is via *social welfare*: the sum of (expected) utilities that players receive[3]. In the search setting, we are mostly interested in the "welfare" of users. Ben Basat et al. [10, 11] made the simplifying assumption that user and publisher utilities are aligned.

Price of anarchy [67, 92] is a useful concept for characterizing a game. It is the ratio between the maximal social welfare that can be attained in a game (not necessarily in an equilibrium) where players collaborate in a selfless manner and the social welfare of the worst equilibrium; i.e., the one with the lowest social welfare. The lower the price of anarchy, the less players' utilities are potentially degraded due to their selfish behavior.

Now, in the ranking games, the ranking function determines the utilities provided to the publishers. Thus, different ranking functions entail different games with potentially different prices of anarchy. Since the true relevance status of documents is assumed to be known as mentioned above [10, 11] — i.e., based on the topical match between the document and the query — it is only natural to assume that documents should be ranked by their relevance status so as to maximize utilities. In fact, this ranking approach is exactly the probability ranking principle (PRP) [88] which was shown to be optimal under some mild conditions in the standard retrieval setting. That is, according to the PRP, documents should be ranked in descending order of their relevance probabilities where these probabilities are estimated using all the information available to the search system. The fundamental question that Ben Basat et al. [10, 11] explored was whether the PRP is optimal in terms of price of anarchy with respect to other ranking approaches.

Ben Basat et al. [10, 11] found that for a game with single topic documents, the PRP is indeed optimal in terms of price of anarchy[4]. However, for multi-topic documents this was not the case anymore. They proposed a stochastic ranking approach which outperformed the PRP in terms of price of anarchy. The sub-optimality of the PRP can be explained using the following toy example.

Suppose a publisher wrote a document d that discusses two topics: A and B, where A is quite common in the corpus and B is unique to d. Further suppose that the publisher is interested in having d highly ranked for queries about topic A. Now, say that d is in fact not highly ranked for A. As a response, the publisher might remove the information about topic B from d to make d more focused, and consequently improve its chances for a better ranking. As a result, the corpus does not contain information about topic B and information needs about B cannot be satisfied. In other words, the topical diversity of the corpus was hurt due to responses to rankings induced by the PRP. Now, Ben Basat et al. [10, 11] suggested a stochastic ranking approach where documents with similar relevance status will be swapped with some probability in the ranking in case the lower ranked one is of higher topical diversity than the higher ranked one. As noted above, this resulted in games with price of anarchy better than that of using the PRP.

Since the PRP essentially serves as the theoretical underpinning of most retrieval methods, whether classical sparse approaches or neural ones [52, 72, 77], its sub-optimality is a significant issue. More generally, the PRP and retrieval methods devised based on it do not account for post-ranking corpus effects which can turn out to be quite harmful in terms of the ability to satisfy information needs.

The findings about post-ranking effects are not only theoretical. Raifer et al. [87] showed using controlled ranking competitions that diversity of content is hurt due to ranking incentives. In addition, they provided a game theoretic approach that explains the empirically observed document modification strategies of publishers regardless of the ranking principle employed. Furthermore, a recent study of historical snapshots of the ClueWeb09 dataset showed that such document modification strategies that lead to reduced topical diversity are potentially also evident on the Web [104].

4.2 Towards a Rigorous Game Theoretic Modeling

As noted above, the framework proposed by Ben Bast et al. [10, 11] is based on assuming that the relevance status of a document for a query is known. This leads us to the first fundamental challenge:

RD1: Devising game theoretic modeling of the competitive search setting where document relevance estimates are used.

This modeling can significantly depart from that of Ben Basat et al. [10, 11]. It can potentially be based on approaches developed for analyzing games with mediators in algorithmic game theory [4, 5, 7, 81, 97]; specifically, those where players have partial information about the world [5]. Recall that the ranking function plays the role of a mediator in the ranking games.

A fundamental game-theoretic perspective that is worth considering is modeling competitive search as a facility location game [3, 18]. Treating the ranking task as a facility location problem, both the query and the documents are modeled as points in a latent space. Documents are ranked by their distance from the query. In the ranking game, publishers select the content to write; for example, the publisher's choice of multiple topics is a strategy profile in the game. The utilities are determined as a function of users' exposure to the content depending on the document's position in the ranking.

[1] This definition corresponds to the state-of-affairs in work on focused retrieval where the percentage of relevant text in a document is the basis for evaluation [44].

[2] We write "expected" since mixed strategies are based on probability distributions.

[3] Social welfare can be computed for any state of a game, not necessarily an equilibrium.

[4] They also showed that if publishers have differential writing qualities for topics, and these are accounted for in the utility/reward mechanism, then even for single topic documents the PRP is not optimal.

Facility location games were discussed for recommendation systems [12]. In this setting, there is a single item recommended rather than a ranked list of documents as in the ranking games. Ben Porat and Tennenholtz [12] showed that using the PRP to recommend a single item even fails to converge to equilibrium; i.e., the PRP is an unstable recommendation principle. They devised alternative stable content selection functions for publishers.

The study of competitive search as a facility location game in which we account for ranking — e.g., publishers' payoffs are determined based on documents' ranks — is fundamental to the understanding of search dynamics, as well as to introduce a novel facility location game setup. This setting will also allow to pose various requirements about the the ranking function, such as diversification [96] and fairness [25] which we discuss below.

Additional fundamental assumption made by Ben Basat et al. [10, 11] was that user and publisher utility are aligned. This is obviously not the case in practical settings. Hence, an important aspect of devising a game theoretic framework is:

RD2: Devising game theory models for ranking that accommodate different user and publisher utilities.

Beyond Relevance. Even in the standard static corpus retrieval setting, relevance is not always the sole criterion by which ranking should be induced. For example, results diversification methods are intended to improve the coverage of query aspects [96]. These methods are often applied as a re-ranking step. As is the case for competitive retrieval settings, the PRP is not optimal when inter-document relevance is considered. An alternative ranking principle was suggested using quantum mechanics concepts [118].

Accounting for inter-document relations, which are the foundation of results diversification approaches, in a game theoretic framework leads to an interesting challenge. Publishers opting to improve the ranking of their documents should now make their documents somewhat different than other documents — either at the surface level or with respect to aspects discussed in the documents. This leads to a novel suite of potential strategies of publishers, departing from the analysis of Basat et al. [10, 11]. In fact, the incomplete knowledge about how other publishers will form their documents leads to a natural Bayesian game setting where players have Bayesian estimates about the "world" in which they operate [56].

Additional foundational example where relevance is not the only criterion by which ranking is induced is fairness, with a specific focus on fairness of exposure of publishers [15, 25, 98, 114]. That is, due to the attention bias of users for top-ranked documents, fairness mechanisms have been suggested to improve the exposure of different publishers with minimal degradation of relevance effectiveness. Some of these mechanisms are based on stochastic ranking approaches [15] and their evaluation is based on the long run effects [34]. The notion of fairness can potentially be accounted for via the definition of utility for publishers. This further strengthens the need to decouple user and publisher utilities.

The results diversification and fairness examples give rise to a more fundamental research direction:

RD3: Devising game theory models that account for multiple ranking criteria with potential inter-publisher effects in the same ranking.

User Perspective and Interactive Retrieval. Corpus dynamics, specifically as a result of ranking incentives, has not been addressed in most previous work. In contrast, the dynamics of search sessions in interactive retrieval has long been studied [9, 55, 66, 91, 93, 101, 110]. Furthermore, there is a growing research interest in conversational search systems with which users interact [43, 113]. Interactive retrieval is a notable setting wherein the PRP, which was designed for a single shot retrieval, does not hold. A ranking principle for interactive retrieval was also devised [42].

The corpus dynamics driven by competing publishers can also potentially affect the queries used by users who respond to induced rankings. For example, if search effectiveness degrades for a query due to the content effects of the publishers' competition, users might start to use somewhat different queries to express the same information needs. This need not be a part of an interactive query session but can rather happen along time. Furthermore, different groups of users might respond differently. A case in point, search personalization can, and should, be affected by the user dynamics entailed by the corpus dynamics.

Integrating all the types of dynamics just mentioned in the same game theoretic framework is an important challenge:

RD4: Modeling simultaneously user dynamics (e.g., in terms of queries posted), system-user dynamics (e.g., in interactive/conversational search) and publishers-corpus dynamics.

Auxiliary Relevance Signals. The competition between publishers is obviously not limited to content. For example, link spam is a long known phenomenon [53, 54]. The underlying incentive is ranking: PageRank scores and alike are used in retrieval methods [19, 73]. Accordingly, there has been much work on addressing hyperlink-based methods that are more robust to spamming [54, 60, 111].

The resultant state-of-affairs is that corpus dynamics due to ranking incentives is manifested in the documents themselves as well as in auxiliary information[5]. Since both types of dynamics affect relevance estimation, an emerging challenge is:

RD5: Modeling simultaneously within-document dynamics and that of auxiliary information.

Concluding Notes. Overall, the importance of game theoretic modeling is providing fundamental theoretical infrastructure for devising (as we discuss in Section 5.1) and contrasting retrieval

[5] Additional type of auxiliary information is anchor text, for example.

principles and approaches in the competitive retrieval setting; e.g., the comparison of the PRP with a stochastic retrieval approach discussed above [10, 11]. In the standard retrieval setting, the ability to formally and theoretically contrast retrieval approaches is highly limited. In past work, diagnostic comparative evaluation of retrieval methods applied to a static corpus was based on properties such as term frequency, inverse document frequency and document length [37].

We note that corpus effects are not limited to topical diversity. The effects can be related to writing style, correctness of information, etc. In Section 7 we discuss a recent work [50] which demonstrated a few types of content effects caused by ranking incentives. These types of effects can be considered in the utility functions used in a game theoretic modeling.

5 THE RANKER PERSPECTIVE

Heretofore we focused on using a game theoretic approach to model games entailed by a choice of a ranking function. We used the price-of-anarchy concept to reason about a game. We now turn to discuss the novel ways ranking functions should be devised in competitive retrieval settings.

5.1 Beyond Myopic Learning-To-Rank

Feature-based learning-to-rank [73] and neural-network-based [51, 52, 72, 77, 108] retrieval methods are based on learning a ranking function using training data composed of documents, queries and relevance judgments. The learning process is guided by optimizing for a loss function which represents either directly or indirectly a relevance effectiveness evaluation measure. None of the loss functions suggested up to date accounts for post-ranking effects, namely, the strategic modification of documents by their publishers who respond to induced rankings.

The state-of-affairs just described, together with the need to account for long-term corpus effects, leads us to the next significant challenge:

RD6: Devising ranking functions that are optimized simultaneously for short-term relevance effectiveness and long-term corpus effects.

By "short term" we mean the retrieved results presented to the user (e.g., search results page); that is, optimizing for relevance effectiveness as is standard using existing loss functions (e.g., AP or NDCG). By "long term" we mean corpus effects that result from publishers' responses to induced rankings. As noted above, these effects can touch on topical coverage and diversity in the corpus, writing style, trustworthiness of content, and more. In quantitative terms, long term effectiveness can be measured by the price of anarchy where utilities, and accordingly social welfare, are defined in terms of the corpus effects we opt to optimize for.

Optimizing simultaneously for short term retrieval effectiveness and for long term price of anarchy, or any other concept that can be used to quantitatively characterize equilibiria in games, is a completely novel research agenda to the best of our knowledge[6]. We propose two major research directions.

The first is devising stochastic ranking mechanisms that allow for simultaneous short term and long term optimization. As already mentioned, in Ben Bast et al.'s work [10, 11], the positions of *known relevant* documents which had similar retrieval scores in a ranked list were flipped with some probability. The goal was to improve the price of anarchy — specifically, in terms of topical coverage. Promoting in ranking documents with relatively high topical diversity can provide their authors an incentive to maintain this diversity or even increase it, thus contributing to the topical diversity in the corpus. Here, our goal and main challenge is to devise realistic stochastic retrieval methods where the relevance status of documents is unknown. To this end, one potential direction is utilizing bounds on pairwise loss functions (or surrogates thereof) [22, 63, 73] in price of anarchy computations. Specifically, one can flip documents in rankings with some probability in case such flips are beneficial to the long-term price of anarchy; at the same time, we should bound, and control for, the short term effectiveness loss using the bounds just mentioned.

The second direction for short-long term optimization of retrieval methods is devising surrogate functions for long-term objectives; specifically, price of anarchy. For example, one can use simulation during the training phase of a retrieval method to estimate the long-term price of anarchy. As a result, a specific ranking can be assigned with a price-of-anarchy estimate which can then be used together with a short-term standard loss function (e.g., AP or NDCG) to yield an optimization criterion. To perform the integration, recently suggested frameworks for learning ranking functions using multiple short-term objectives can be used [24]. Additional direction is using bi-level optimization mechanisms [32]. In the bi-level optimization setting one optimization problem is embedded in another. Often, the output of one problem serves as the input of the other. Rather than applying serial optimization (i.e., optimizing the first problem and then the second), one can use bi-level procedures for simultaneous optimization.

Obviously, simulating long-term effects during the training phase is difficult due to the uncertainty about publishers' responses to induced rankings. Hence, one can set as a goal to devise "publisher-response" estimated models that can be used for the simulation; e.g., a model that assumes addition of query terms to a document as a means to promoting it in future rankings.

5.2 Effectiveness of Content-Based Relevance Estimates under Strategic Document Manipulations

Content-based relevance estimates play a key role in feature-based learning-to-rank methods [73] and neural methods [52, 72, 77]. Standard estimates, which utilize term frequency information (TF), are vulnerable to a simple strategic manipulation: stuffing terms from queries of interest in the document. In feature-based ranking functions, this vulnerability is compensated for by using spam

[6] Ben Basat et al. [10, 11] optimized for long term price of anarchy. Ghosh and McAfee [45, 46] compared mechanisms to enhance content quality of content created by ad hoc contributors. Neither of these lines of work addressed the simultaneous short-long term optimization challenge we refer to here.

classifiers and various other document quality measures [13]. However, white-hat document modification strategies as light keyword stuffing need not hurt the quality of a document or turn it into spam.

There is a growing body of work on adversarial attacks on neural-networks-based approaches (e.g., classifiers) [35, 58, 71, 84, 100, 107]. This line of work drove forward work on improving the robustness of machine-learning based approaches to adversarial manipulations; e.g., [57, 62]. There is very little work we are aware of on studying the effectiveness of neural rankers under adversarial attacks [106]. We are also not aware of work on improving their effectiveness in competitive retrieval settings. We note that some neural retrieval methods integrate lexical-matching modules [78]; hence, they become vulnerable at the lexical level. Furthermore, there is recent work on adversarial attacks on BERT [71]. BERT (and other pre-trained language models) are the basis of highly effective ranking models which rely on representation learning [72].

Hence, one direction which we view as highly important is devising content-based relevance estimates that are effective under strategic content manipulations. In what follows we describe a few avenues that can be pursued to this end.

Classical Rankers. Okapi BM25 [90] is one of the most effective classical ("sparse") content-based relevance estimates that till this day serves as a reference comparison to neural methods. It is based on the assumption that the occurrences of *elite* and *non-elite* terms in documents are both distributed poisson with different means. The TF-based component of Okapi BM25 is an approximation to a two-poisson mixture model [89].

In the competitive retrieval setting, occurrences of elite terms can be the result of strategic document manipulations. One way to adapt Okapi BM25 to a competitive retrieval setting is to assume a three component poisson mixture model: terms are either (i) not elite, (ii) elite but result from strategic manipulations, or (iii) elite without originating from such manipulations. The main challenge here is to estimate the extent of query terms occurrence that can be attributed to "pure eliteness" and that which should be attributed to strategic manipulations.

The axiomatic framework for ad hoc retrieval [38] is an example of theoretical grounds for classical retrieval methods. The framework enables to contrast existing retrieval methods and to devise novel ones. One of the axioms is that increase of query-term occurrence in a document should not decrease, and should often increase, the document retrieval score. However, increased query-term occurrence in a competitive retrieval setting might be due to strategic document modifications. Hence, re-visiting the axiomatic framework for competitive retrieval setting is an intriguing research avenue.

Okapi BM25 and the axiomatic framework are two examples of retrieval approaches out of many other classical methods that need to be re-considered for competitive retrieval settings:

RD7: Adapting classical content-based retrieval methods to competitive retrieval settings with strategic document manipulations.

Neural Rankers. Some neural-based approaches utilize lexical similarities between a query and documents, others use representation learning (e.g., by utilizing pre-trained language models [72]) to infer semantic query-document similarities, and some integrate the two paradigms [52, 77]. As noted above, there is an increasing body of work in the machine learning and natural language processing communities on improving the effectiveness of neural-network-based approaches (and others) with respect to adversarial examples; e.g., [57, 62, 116]. Adapting some of these approaches which are pointwise — as they operate mainly in classification tasks — to the ranking setting, where the relative ordering of documents is important rather than their pointwise relevance estimates, might be a first step towards improving these methods. More generally, the next significant research challenge we point to is:

RD8: Improving the effectiveness of neural retrieval methods under strategic content manipulations.

Ranking Robustness. Goren et al. [48] noted that one of the potential consequences of ranking competitions is instability of rankings. That is, that users will browse search results pages that rapidly change due to potentially small document changes that are due to ranking incentives. They defined different notions of ranking robustness, that is, the changes in the ranking of a document set as a result of document modifications[7]. They showed that the stronger the regularization of a learned ranking function, the more robust the rankings it induces. A major question left open as a result of the study of Goren et al. [48] is:

RD9: Balancing search effectiveness and the robustness of induced rankings.

That is, the goal is to devise learning-to-rank functions that account for both aspects. At one extreme, fixed rankings regardless of document changes are highly robust but search effectiveness can be significantly hurt. At the other extreme, highly effective rankings with very similar document retrieval scores can be quite unstable with respect to small document modifications.

6 THE PUBLISHER PERSPECTIVE

A foundational aspect of the competitive retrieval setting is the responses of publishers to rankings, specifically, by applying strategic document manipulations. These are intended to increase the chances of a document to be more highly ranked in future rankings. A basic fundamental research challenge is then (cf., [49]):

RD10: Devising methods for automatic manipulation of documents that will potentially improve its future ranking for queries of interest and will not hurt its quality (e.g., coherence).

We should distinguish this challenge from that of spamming or more generally *black hat* search engine optimization [53]. Goren et

[7] This notion should be differentiated from various notions of performance robustness of a retrieval method; e.g., its performance decay due to strategic document manipulations.

al. [49] noted that there are three desiderata for white hat (i.e., legitimate) automatic manipulations: (i) keeping the modified document faithful to the original document, (ii) maintaining the quality of the document, (iii) improving its chances to be more highly ranked in future rankings.

Why should the research community care about legitimate automatic document modification, namely, white hat SEO? Goren et al. [49] pointed to two main reasons. First, given the dramatic advanced in language modeling capabilities, and accordingly, approaches to text generation, the search echosystem will be composed of more and more texts that are automatically generated with the goal to improve ranking. Those who will not have at their disposal automatic modification capabilities will not be competitive; i.e., this is a publishers fairness issue. The second reason is the need to produce synthetic data (i.e., automatically generated texts) for evaluation in competitive retrieval settings. Obtaining real world data about ranking competitions is an extremely hard task. It requires access to logs of commercial search engines, and even then, isolating specific phenomenon is an extremely hard task as we discuss in Section 8. A third important reaon, mentioned in Section 1, is to push forward work on improving the effectiveness of retrieval methods under strategic document manipulations. In the crypto community, for example, attacks on encryption methods helped to push progress on devising more effective methods.

Goren et al.'s approach [49] for automatic document manipulation was based on replacing a passage in a document with a passage from another document. The criteria for selecting which passage to replace, and which passage will replace it, were based on estimated rank promotion by the undisclosed ranking function and document coherence. Using modern language generation techniques for automatic document modification is obviously a next step. This task could be viewed as paraphrasing where one opts to keep the document coherent and of high quality and at the same time change it so that it is potentially more highly ranked in future rankings.

7 SOCIETAL EFFECTS

Raifer et al. [87] analyzed using game theoretic modeling a potential document modification strategy of publishers to induced rankings. They provided formal support to the strategy of mimicking content in documents that were highly ranked in the past for the query at hand. Indeed, induced rankings are the only signal about the undisclosed ranking function. Controlled ranking competitions between students, which we refer to in more detail below, provided empirical support to the prevalence of this strategy [87]. Later on, Vasilisky et al. [104] analyzed historical snapshots of TREC's ClueWeb09 collection, and found that documents highly ranked for the queries were indeed becoming more and more similar to each other along time.

All the findings just stated provide further support to Ben Basat et al.'s [10, 11] theoretical finding that using the PRP results in decreased (topical) diversity. While this has a negative effect on the ability to satisfy future information needs, as already discussed, there are even more worrisome phenomena that arise and which we discuss next.

The mimicking strategy is an example of the well known *herding* effect studied in depth in the economics literature [8, 16, 99]. Goren et al. [50] showed that herding (mimicking) can be along various dimensions. For a proof of concept, they organized ranking competitions between students and manually positioned at the first rank documents that manifested a desired effect. The first effect they studied was with respect to query aspects (as defined for TREC topics) discussed in documents. They found that positioning at the first rank a document that discusses one aspect of a query but not the others resulted in dynamics that shifted documents' content to this aspect; i.e., via document modifications applied by the students. The second type of document they positioned at the first rank was one that contained the query terms but was not relevant to the query. The resultant dynamics along time was that fewer and fewer relevant documents were found in the corpus. The third type of effect was document length: the result of positioning a short document at the first rank was that students shortened the lengths of their documents along time. The fourth effect was the inclusion, or lack thereof, of query terms in a document. Positioning at the first rank a document which did not contain the query terms but was relevant to the query resulted in corpus dynamics with diminishing number of documents containing the query terms.

The findings just described with regard to herding effects also attest to the ability of publishers to "distill" potential biases of a ranking function as manifested in highly ranked documents. As Goren et al. [50] noted, there is a fundamental potential issue with this state-of-affairs. If a publisher opts to promote in the corpus some negative effect (e.g., improper use of language, hate speech, fake news, etc.), the search engine becomes a potential platform to drive this effect. That is, if the safety mechanisms of the search engine fail to identify issues with the document, and it is highly ranked, then there is a potential effect on the corpus. Thus, while most concern thus far with respect to biases of ranking functions have been with regard to users of the search engine browsing the search results page, or fairness to publishers, there is an additional potentially pervasive issue: corpus effects. This reality gives rise to two strategic research directions which have not been explored thus far for ad hoc retrieval. The first is:

RD11: Analyzing ranking-driven herding phenomena in a given corpus.

We note that while content changes in the Web have been studied [85, 86, 95], they were not analyzed from a ranking-driven perspective. The second research direction we point to is:

RD12: Analyzing ranking functions for potential biases that can lead to herding effects.

This research direction is intended to address concerns conceptually similar to those addressed in the recently published U.S. bill on algorithmic accountability[8]. Algorithmic accountability in the bill refers to algorithm-based decisions that affect people (e.g., in health, employment, etc.) While as already mentioned, users affected by search results they are exposed to, they are affected in the long run by trends in the corpus which can be due to herding effects.

[8] H.R.6580 Algorithmic Accountability Act of 2022.

Finally, we note that effects on content due to algorithmic decisions and herding phenomenon have already been reported in the past. For example, in the Facebook experiment, the sentiment expressed in posts was affected by that in promoted content [68]. However, there has not been work, to the best of our knowledge, on analyzing and addressing this phenomena in the ad hoc retrieval setting.

8 EMPIRICAL ANALYSIS

Studying corpus dynamics driven by ranking incentives is a very difficult challenge. Competitive search settings are served by large scale commercial search engines (e.g., Google, Bing, Baidu, Yandex, etc.). Hence, to attribute corpus changes to rankings one would need access to proprietary data, namely, search logs containing ranking information. But even with access to search logs, the task remains highly difficult: corpus dynamics can be due to various reasons which are not necessarily related to ranking.

Additional challenge of empirical analysis in competitive search settings is the evaluation of novel ranking methods. Since publishers might respond to induced rankings, the standard approach of using a fixed corpus falls short as post-ranking effects are not measured. Hence, evaluation calls for a "live" setting where publishers observe rankings and respond.

Raifer et al. [87] and Goren et al. [49, 50] addressed the challenges just described by organizing small scale ranking competitions. Students in courses served as publishers and were assigned to queries. Their goal was to write short plain text documents that would be highly ranked for the queries. The students went through a few rounds of the competition in which they were shown a ranking induced by an undisclosed ranking function. They could respond to the rankings by modifying their documents.

The corpora that resulted from Raifer et al.'s [87] and Goren et al.'s [49, 50] ranking competitions allowed to study, in a post hoc manner, the corpus dynamics. As already mentioned, there was clear evidence for herding phenomena [50, 87]. Obviously, running such a ranking competition, even if small scale, to evaluate each novel retrieval method — specifically, its effects on the corpus – is challenging. Another important aspect of using controlled competitions is the incentive. Raifer et al. [87] increased the incentives at some point of the competitions due to relatively low level of dynamics.

Additional potential challenge with using controlled small scale competitions is the extent to which the findings transfer to large scale and evolved competitive settings. As described above, Vasilisky et al. [104] addressed this question and found that documents highly ranked for a query in past snapshots of the ClueWeb09 corpus were becoming more and more similar to each other along time. This phenomenon was aligned with Reifer et al.'s and Goren et al.'s findings [50, 87]. Yet, Vasilisky et al. noted that their findings could potentially be attributed to reasons other than responses to rankings for the specific queries they used.

Given the above, the following direction becomes quite important for research on competitive retrieval:

RD13: Devising sustainable large scale and evolved ranking competitions.

One of the ways to ameliorate the need for human publishers participating in ranking competitions is devising automatic approaches for document manipulations as discussed in Section 6. Goren et al. [50] used a simple automatic document modification approach as a bot in ranking competitions held between students. They found that the bot produced documents that were more highly ranked on average than those of students, and that were judged to be of high quality. Using such bots to create content and studying the resulting ranking-incentivized corpus dynamics seems a promising direction towards evaluation in settings with much automatically created content. Nevertheless, ensuring that content created by bots is similar in characteristics to that created by humans, and defining content modification strategies of bots to follow those of humans, is still a highly challenging future direction. Success with this direction will result in improved ability to perform offline simulation of ranking competitions.

9 SUMMARY

Ranking incentives of publishers and the entailed corpus dynamics constitute the basis of a competitive search echosystem. Although the largest-scale competitive setting in the data science realm is Web search, competitive search has not attracted much research attention in the information retrieval community except for work on spam and low quality documents [13, 27]. Perhaps the main reason is the difficulty of evaluation. As a case in point, novel ranking functions cannot be properly evaluated using a static corpus as corpus dynamics is not accounted for. Rather, a "live" setting where publishers respond to induced rankings is called for.

From a scientific point of view, the theoretical underpinning of the core task in information retrieval, ranking, breaks down in competitive settings; that is, the probability ranking principle [10, 11] (PRP) is not optimal as it leads to reduced topical diversity. More generally, retrieval functions devised throughout the years, the vast majority of which adhere to the PRP, do not account for post-ranking effects on the corpus that are due to ranking incentives of publishers. At the same time, in other research communities, there is a growing body of work on adversarial aspects (e.g., [58, 62, 71, 84, 100, 107]).

Game theory provides effective grounds to model some of the dynamics of the competitive retrieval setting [10, 11]. We argued that a rigorous game theoretical framework which will allow to formally contrast retrieval approaches and devise novel ones is called for. Furthermore, such a framework should not be confined to relevance ranking for a single query, but should also account for diversification, fairness, interactive modes of retrieval, changes of user queries along time as responses to corpus changes, personalization aspects in search, and other aspects of retrieval. (RD1-RD5)

We argued for a novel way of learning ranking functions in the competitive search setting (RD6). Rather than optimizing for a myopic loss function which is the standard approach, one should also account for long-term corpus effects. Furthermore, current relevance estimates, whether classical or modern, should be adapted to a realm where documents are manipulated for improved ranking (RD7-RD9)

We then moved to discuss the publisher perspective (RD10). We discussed the importance of devising approaches to automatic document manipulations that potentially help to improve the document ranking and at the same time maintain its quality. An important motivation for engaging in this line of work is the ability to improve the effectiveness of retrieval methods that face strategic document manipulations as is the case in the crypto community where published attacks on encryption algorithms drive forward the development of improved algorithms.

The competitive search setting also entails societal effects. Due to the herding phenomenon [50], publishers can potentially exploit the search engine so as to drive forward desired content effects on the corpus. Thus, while many of the concerns thus far with respect to search engines were about effects on its users, and fairness to publishers, we argue that concerns about corpus effects should also be addressed (RD11-RD12).

Empirically analyzing retrieval in competitive retrieval settings is a difficult challenge due to the complex dynamics of the corpus which is not entirely driven by ranking incentives. Findings that emerged from controlled ranking competitions held between students [49, 50, 87] demonstrated the considerable potential of such evaluation. Yet, increasing the scale of the evaluation, allowing for offline simulated evaluation (using automatically modified documents), and accounting for auxiliary document information (e.g., hyperlinks) in addition to content are still all research avenues that should be pursued in our opinion (RD13).

Acknowledgments. We thank the reviewers for their comments. The work by Moshe Tennenholtz was supported by funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement 740435).

REFERENCES

[1] 2005–2009. *AIRWeb — International Workshop on Adversarial Information Retrieval on the Web.*

[2] Gianni Amati and C. J. van Rijsbergen. 2002. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Trans. Inf. Syst.* 20, 4 (2002), 357–389.

[3] Simon P. Anderson, Andre de Palma, and Jacques-Francois Thisse. 1992. *Discrete Choice Theory of Product Differentiation.* MIT press.

[4] Itai Ashlagi, Dov Monderer, and Moshe Tennenholtz. 2008. On the Value of Correlation. *J. Artif. Intell. Res. (JAIR)* 33 (2008), 575–613.

[5] Itai Ashlagi, Dov Monderer, and Moshe Tennenholtz. 2009. Mediators in position auctions. *Games and Economic Behavior* 67, 1 (2009), 2–21. https://doi.org/10.1016/j.geb.2008.11.005

[6] Susan Athey and Glenn Ellison. 2011. Position Auctions with Consumer Search. *The Quarterly Journal of Economics* 126, 3 (2011).

[7] Robert Aumann. 1974. Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics* 1 (1974), 67–96.

[8] Banerjee. 1992. A simple model of herd behavior. *The Quarterly Journal of Economics* 107 (1992), 797–817.

[9] Nicholas J. Belkin, Colleen Cool, Diane Kelly, S-J Lin, SY Park, J Perez-Carballo, and C Sikora. 2001. Iterative exploration, design and evaluation of support for query reformulation in interactive information retrieval. *Information Processing & Management* 37, 3 (2001), 403–434.

[10] Ran Ben-Basat, Moshe Tennenholtz, and Oren Kurland. 2015. The Probability Ranking Principle is Not Optimal in Adversarial Retrieval Settings. In *Proceedings of ICTIR.* 51–60.

[11] Ran Ben-Basat, Moshe Tennenholtz, and Oren Kurland. 2017. A Game Theoretic Analysis of the Adversarial Retrieval Setting. *J. Artif. Intell. Res.* 60 (2017), 1127–1164.

[12] Omer Ben-Porat and Moshe Tennenholtz. 2018. A Game-Theoretic Approach to Recommendation Systems with Strategic Content Providers. In *Proc. of NIPS.* 1110–1120.

[13] Michael Bendersky, W. Bruce Croft, and Yanlei Diao. 2011. Quality-biased ranking of Web documents. In *Proceedings of WSDM.* 95–104.

[14] Ron Berman and Zsolt Katona. 2013. The Role of Search Engine Optimization in Search Marketing. *Mark. Sci.* 32, 4 (2013), 644–651.

[15] Asia J. Biega, Krishna P. Gummadi, and Gerhard Weikum. 2018. Equity of Attention: Amortizing Individual Fairness in Rankings. In *Proceedings of SIGIR.* 405–414.

[16] S. Bikhchandani, D. Hirshleifer, and I. Welch. 1992. A theory of fads, fashion, custom and cultural change as information cascade. *The Journal of Political Economy* 100 (1992), 992–1026.

[17] István Bíró, Jácint Szabó, and András A. Benczúr. 2008. Latent dirichlet allocation in web spam filtering. In *Proceedings of AIRWeb 2008, Fourth International Workshop on Adversarial Information Retrieval on the Web.* 29–32.

[18] S. Brenner. 2010. Location (hotelling) games and applications. Wiley Encyclopedia of Operations Research and Management Science.

[19] Sergey Brin and Lawrence Page. 1998. The Anatomy of a Large-Scale Hypertextual Web Search Engine. In *Proceedings of WWW.* 107–117.

[20] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165

[21] Chris Buckley, Gerard Salton, James Allan, and Amit Singhal. 1994. Automatic query expansion using SMART: TREC3. In *Proceedings of TREC.* 69–80.

[22] Christopher J. C. Burges. 2010. *From RankNet to LambdaRank to LambdaMart: An overview.* Technical Report. Microsoft.

[23] James P. Callan. 1994. Passage-level Evidence in Document Retrieval. In *Proceedings of SIGIR.* 302–310.

[24] David Carmel, Elad Haramaty, Arnon Lazerson, and Liane Lewin-Eytan. 2020. Multi-Objective Ranking Optimization for Product Search Using Stochastic Label Aggregation. In *Proc. of the WebConf.* 373–383.

[25] Carlos Castillo. 2018. Fairness and Transparency in Ranking. *SIGIR Forum* 52, 2 (2018), 64–71.

[26] Carlos Castillo, Claudio Corsi, Debora Donato, Paolo Ferragina, and Aristides Gionis. 2008. Query-log mining for detecting spam. In *Proceedings of AIRWeb, Fourth International Workshop on Adversarial Information Retrieval on the Web.* 17–20.

[27] Carlos Castillo and Brian D. Davison. 2010. Adversarial Web Search. *Foundations and Trends in Information Retrieval* 4, 5 (2010), 377–486.

[28] Carlos Castillo, Debora Donato, Luca Becchetti, Paolo Boldi, Stefano Leonardi, Massimo Santini, and Sebastiano Vigna. 2006. A reference collection for web spam. *SIGIR Forum* 40, 2 (2006), 11–24.

[29] Carlos Castillo, Debora Donato, Aristides Gionis, Vanessa Murdock, and Fabrizio Silvestri. 2007. Know your neighbors: web spam detection using the web topology. In *Proceedings of SIGIR.* 423–430.

[30] Gordon V. Cormack, Mark D. Smucker, and Charles L. A. Clarke. 2011. Efficient and effective spam filtering and re-ranking for large web datasets. *Informaltiom Retrieval Journal* 14, 5 (2011), 441–465.

[31] W. Bruce Croft and John Lafferty (Eds.). 2003. *Language Modeling for Information Retrieval.* Number 13 in Information Retrieval Book Series. Kluwer.

[32] Stephan Dempe. 2002. *Foundations of Bilevel Programming.* Springer.

[33] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR* abs/1810.04805 (2018).

[34] Fernando Diaz, Bhaskar Mitra, Michael D. Ekstrand, Asia J. Biega, and Ben Carterette. 2020. Evaluating Stochastic Rankings with Expected Exposure. In *Proceedings of CIKM.* 275–284.

[35] Yinpeng Dong, Hang Su, Baoyuan Wu, Zhifeng Li, Wei Liu, Tong Zhang, and Jun Zhu. 2019. Efficient Decision-Based Black-Box Adversarial Attacks on Face Recognition. In *Proceedings of CVPR.* 7714–7722.

[36] Kfir Eliaz and Ran Spiegler. 2011. A simple model of search engine pricing. *The Economic Journal* 121, 556 (2011), F329–F339.

[37] Hui Fang, Tao Tao, and ChengXiang Zhai. 2011. Diagnostic Evaluation of Information Retrieval Models. *ACM Trans. Inf. Syst.* 29, 2 (2011), 7:1–7:42.

[38] Hui Fang and ChengXiang Zhai. 2005. An exploration of axiomatic approaches to information retrieval. In *Proceedings of SIGIR.* 480–487.

[39] Dennis Fetterly, Mark Manasse, and Marc Najork. 2004. Spam, Damn Spam, and Statistics: Using Statistical Analysis to Locate Spam Web Pages. In *Proceedings of WebDB.* 1–6.

[40] Andrey Fradkin. 2017. Search, Matching, and the Role of Digital Marketplace Design in Enabling Trade: Evidence from Airbnb. Available at SSRN: https://ssrn.com/abstract=2939084 or http://dx.doi.org/10.2139/ssrn.2939084.

[41] Andrey Fradkin. 2019. A Simulation Approach to Designing Digital Matching Platforms. Available at SSRN: https://ssrn.com/abstract=3320080 or http://dx.doi.org/10.2139/ssrn.3320080.

[42] Norbert Fuhr. 2008. A probability ranking principle for interactive information retrieval. *Information Retrieval* 11, 3 (2008), 251–265.

[43] Jianfeng Gao, Chenyan Xiong, Paul Bennett, and Nick Craswell. 2022. Neural Approaches to Conversational Information Retrieval. *CoRR* abs/2201.05176 (2022).

[44] Shlomo Geva, Jaap Kamps, and Ralf Schenkel (Eds.). 2012. *Focused Retrieval of Content and Structure, 10th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX.* Lecture Notes in Computer Science, Vol. 7424.

[45] Arpita Ghosh and R. Preston McAfee. 2011. Incentivizing high-quality user-generated content. In *Proceedings of WWW.* 137–146.

[46] Arpita Ghosh and R. Preston McAfee. 2012. Crowdsourcing with endogenous entry. In *Proceedings of WWW.* 999–1008.

[47] Yu Gong, Xusheng Luo, Kenny Q. Zhu, Wenwu Ou, Zhao Li, and Lu Duan. 2019. Automatic Generation of Chinese Short Product Titles for Mobile Display. In *Proc. of AAAI.*

[48] Gregory Goren, Oren Kurland, Moshe Tennenholtz, and Fiana Raiber. 2018. Ranking Robustness Under Adversarial Document Manipulations. In *Proceedings of SIGIR.* 395–404.

[49] Gregory Goren, Oren Kurland, Moshe Tennenholtz, and Fiana Raiber. 2020. Ranking-Incentivized Quality Preserving Content Modification. In *Proceedings of SIGIR.* 259–268.

[50] Gregory Goren, Oren Kurland, Moshe Tennenholtz, and Fiana Raiber. 2021. Driving the Herd: Search Engines as Content Influencers. *CoRR* abs/2110.11166 (2021).

[51] Jiafeng Guo, Yixing Fan, Qingyao Ai, and W. Bruce Croft. 2016. A Deep Relevance Matching Model for Ad-hoc Retrieval. In *Proceedings of CIKM.* 55–64.

[52] Jiafeng Guo, Yixing Fan, Liang Pang, Liu Yang, Qingyao Ai, Hamed Zamani, Chen Wu, W. Bruce Croft, and Xueqi Cheng. 2020. A Deep Look into neural ranking models for information retrieval. *Inf. Process. Manag.* 57, 6 (2020).

[53] Zoltán Gyöngyi and Hector Garcia-Molina. 2005. Web Spam Taxonomy. In *Proc. of AIRWeb 2005.* 39–47.

[54] Zoltán Gyöngyi, Hector Garcia-Molina, and Jan O. Pedersen. 2004. Combating Web Spam with TrustRank. In *Proceedgins of VLDB.* 576–587.

[55] Donna Harman. 1988. Towards interactive query expansion. In *Proc. of SIGIR.* 321–331.

[56] John C. Harsanyi. 1967. Games with Incomplete Information Played by "Bayesian" Players, I–III Part I. The Basic Model. *Management Science* 14, 3 (1967).

[57] Warren He, James Wei, Xinyun Chen, Nicholas Carlini, and Dawn Song. 2017. Adversarial Example Defenses: Ensembles of Weak Defenses are not Strong. *CoRR* abs/1706.04701 (2017).

[58] Sandy Huang, Nicolas Papernot, Ian J. Goodfellow, Yan Duan, and Pieter Abbeel. 2017. Adversarial Attacks on Neural Network Policies. In *Proc. of ICLR.*

[59] Peter Izsak, Fiana Raiber, Oren Kurland, and Moshe Tennenholtz. 2014. The search duel: a response to a strong ranker. In *Proceedings of SIGIR.* 919–922.

[60] Jacob, Olivier Chapelle, and Carlos Castillo. 2008. Web spam identification through content and hyperlinks. In *Proceedings of AIRWeb 2008, Fourth International Workshop on Adversarial Information Retrieval on the Web.* 41–44.

[61] N. Jardine and C. J. van Rijsbergen. 1971. The use of hierarchic clustering in information retrieval. *Information Storage and Retrieval* 7, 5 (1971), 217–240.

[62] Robin Jia, Aditi Raghunathan, Kerem Göksel, and Percy Liang. 2019. Certified Robustness to Adversarial Word Substitutions. In *Proceedings of EMNLP-IJCNLP.* 4127–4140.

[63] Thorsten Joachims. 2002. Optimizing Search Engines Using Clickthrough Data. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* 133–142.

[64] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. 2005. Accurately interpreting clickthrough data as implicit feedback. In *Proc. of SIGIR.* 154–161.

[65] Timothy Jones, Ramesh S. Sankaranarayana, David Hawking, and Nick Craswell. 2009. Nullification test collections for web spam and SEO. In *Proceedings of AIRWeb.* 53–60.

[66] Diane Kelly and Xin Fu. 2006. Elicitation of term relevance feedback: an investigation of term source and context. In *Proc. of SIGIR.* 453–460.

[67] E. Koutsoupias and C. Papadimitriou. 1999. Worst-Case Equilibria. In *Proc. of STACS.*

[68] Adam D. I. Kramer, Jamie Elizabeth Guillory, and Jeffrey T. Hancock. 2014. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences of the United States of America* 111 24 (2014), 8788–90.

[69] Oren Kurland and Lillian Lee. 2004. Corpus structure, language models, and ad hoc information retrieval. In *Proceedings of SIGIR.* 194–201.

[70] Victor Lavrenko and W. Bruce Croft. 2001. Relevance-Based Language Models. In *Proceedings of SIGIR.* 120–127.

[71] Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. 2020. BERT-ATTACK: Adversarial Attack Against BERT Using BERT. In *Proc. of EMNLP.* 6193–6202.

[72] Jimmy Lin, Rodrigo Nogueira, and Andrew Yates. 2020. Pretrained Transformers for Text Ranking: BERT and Beyond. *CoRR* abs/2010.06467 (2020).

[73] Tie-Yan Liu. 2011. *Learning to Rank for Information Retrieval.* Springer. I–XVII, 1–285 pages.

[74] Xiaoyong Liu and W. Bruce Croft. 2002. Passage retrieval based on language models. In *Proceedings of CIKM.* 375–382.

[75] Xiaoyong Liu and W. Bruce Croft. 2004. Cluster-Based Retrieval Using Language Models. In *Proceedings of SIGIR.* 186–193.

[76] Rishabh Mehrotra and Benjamin A. Carterette. 2019. Recommendations in a marketplace. In *Proceedings of RecSys*, Toine Bogers, Alan Said, Peter Brusilovsky, and Domonkos Tikk (Eds.). 580–581.

[77] Bhaskar Mitra and Nick Craswell. 2018. An Introduction to Neural Information Retrieval. *Foundations and Trends in Information Retrieval* 13, 1 (2018), 1–126.

[78] Bhaskar Mitra, Fernando Diaz, and Nick Craswell. 2017. Learning to Match using Local and Distributed Representations of Text for Web Search. In *Proceedings of WWW.* 1291–1299.

[79] Mandar Mitra, Amit Singhal, and Chris Buckley. 1998. Improving Automatic Query Expansion. In *Proceedings of SIGIR.* 206–214.

[80] Martin Mladenov, Elliot Creager, Omer Ben-Porat, Kevin Swersky, Richard S. Zemel, and Craig Boutilier. 2020. Optimizing Long-term Social Welfare in Recommender Systems: A Constrained Matching Approach. In *Proceedings of ICML.* 6987–6998.

[81] Dov Monderer and Moshe Tennenholtz. 2004. K-Implementation. *J. Artif. Intell. Res. (JAIR)* 21 (2004), 37–62.

[82] John F. Nash. 1950. Equilibrium Points in N-Person Games. *Proc. of the National Academy of Sciences of the United States of America* 36.1 (1950), 48–49.

[83] Alexandros Ntoulas, Marc Najork, Mark Manasse, and Dennis Fetterly. 2006. Detecting spam web pages through content analysis. In *Proceedings of WWW.* 83–92.

[84] Nicolas Papernot, Patrick D. McDaniel, Ian J. Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. 2017. Practical Black-Box Attacks against Machine Learning. In *Proc. of AsiaCCS.* 506–519.

[85] Kira Radinsky and Paul N. Bennett. 2013. Predicting content change on the web. In *Proceedings of WSDM.* 415–424.

[86] Kira Radinsky, Fernando Diaz, Susan T. Dumais, Milad Shokouhi, Anlei Dong, and Yi Chang. 2013. Temporal web dynamics and its application to information retrieval. In *Proceedings of WSDM.* 781–782.

[87] Nimrod Raifer, Fiana Raiber, Moshe Tennenholtz, and Oren Kurland. 2017. Information Retrieval Meets Game Theory: The Ranking Competition Between Documents' Authors. In *Proceedings of SIGIR.* 465–474.

[88] Stephen E. Robertson. 1977. The Probability Ranking Principle in IR. *Journal of Documentation* (1977), 294–304. Reprinted in K. Sparck Jones and P. Willett (eds), *Readings in Information Retrieval*, pp. 281–286, 1997.

[89] Stephen E. Robertson and Steve Walker. 1994. Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In *Proceedings of SIGIR.* 232–241.

[90] Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at TREC-3. In *Proceedings of TREC.*

[91] Joseph John Rocchio. 1971. Relevance Feedback in Information Retrieval. In *The SMART Retrieval System: Experiments in Automatic Document Processing*, Gerard Salton (Ed.). Prentice Hall, 313–323.

[92] T. Roughgarden and E. Tardos. 2002. How bad is selfish routing? *J. ACM* 49, 2 (April 2002), 236–259.

[93] Gerard Salton and Chris Buckley. 1990. Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science (JASIS)* 41, 4 (1990), 288–297.

[94] Jerard Salton, Anita Wong, and Chung Shu Yang. 1975. A vector space model for automatic indexing. *Commun. ACM* 18, 11 (1975), 613–620.

[95] Aécio S. R. Santos, Bruno Pasini, and Juliana Freire. 2016. A First Study on Temporal Dynamics of Topics on the Web. In *Proceedings of WWW.* 849–854.

[96] Rodrygo L. T. Santos, Craig MacDonald, and Iadh Ounis. 2015. Search Result Diversification. *Foundations and Trends in Information Retrieval* 9, 1 (2015), 1–90.

[97] Y. Shoham and M. Tennenholtz. 1995. Social Laws for Artificial Agent Societies: Off-line Design. *Artificial Intelligence* 73 (1995).

[98] Ashudeep Singh and Thorsten Joachims. 2018. Fairness of Exposure in Rankings. In *Proceedings of SIGKDD.* 2219–2228.

[99] L. Smith and P. Sorensen. 2000. Pathalogical outcomes of observational learning. *Econometrica* 68 (2000), 371–398.

[100] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. In *Proc. of ICLR.*

[101] Bin Tan, Atulya Velivelli, Hui Fang, and ChengXiang Zhai. 2007. Term feedback for information retrieval with language models. In *Proc. of SIGIR.* 263–270.

[102] Moshe Tennenholtz and Oren Kurland. 2019. Rethinking search engines and recommendation systems: a game theoretic perspective. *Commun. ACM* 62, 12 (2019), 66–75.

[103] Florian Tramér, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. 2020. Ensemble Adversarial Training: Attacks and Defenses. arXiv:1705.07204 [stat.ML]

[104] Ziv Vasilisky, Moshe Tennenholtz, and Oren Kurland. 2020. Studying Ranking-Incentivized Web Dynamics. In *Proceedings of SIGIR*. 2093–2096.

[105] Ellen M. Voorhees. 1985. The cluster hypothesis revisited. In *Proceedings of SIGIR*. 188–196.

[106] Chen Wu, Ruqing Zhang, Jiafeng Guo, Yixing Fan, and Xueqi Cheng. 2021. Are Neural Ranking Models Robust?

[107] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan L. Yuille. 2017. Adversarial Examples for Semantic Segmentation and Object Detection. In *Proc. of ICCV 2017*. 1378–1387.

[108] Chenyan Xiong, Zhuyun Dai, Jamie Callan, Zhiyuan Liu, and Russell Power. 2017. End-to-End Neural Ad-hoc Ranking with Kernel Pooling. In *Proc. of SIGIR*. 55–64.

[109] Jinxi Xu and W. Bruce Croft. 1996. Query Expansion using Local and Global Document Analysis. In *Proceedings of SIGIR*. 4–11.

[110] Grace Hui Yang, Marc Sloan, and Jun Wang. 2016. *Dynamic Information Retrieval Modeling*. Morgan & Claypool Publishers.

[111] Haixuan Yang, Irwin King, and Michael R. Lyu. 2007. DiffusionRank: a possible penicillin for web spamming. In *Proceedings of SIGIR*. 431–438.

[112] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In *Proc. of NeurIPS*. 5754–5764.

[113] Hamed Zamani, Johanne R. Trippas, Jeff Dalton, and Filip Radlinski. 2022. Conversational Information Seeking. *CoRR* abs/2201.08808 (2022).

[114] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo A. Baeza-Yates. 2017. FA*IR: A Fair Top-k Ranking Algorithm. In *Proceedings of CIKM*. 1569–1578.

[115] Chengxiang Zhai. 2021. Interactive Information Retrieval: Models, Algorithms, and Evaluation. In *Proceedings of SIGIR*. 2662–2665.

[116] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. 2019. Theoretically Principled Trade-off between Robustness and Accuracy. In *Proceedings of ICML*. 7472–7482.

[117] Jianguo Zhang, Pengcheng Zou, Zhao Li, Yao Wan, Xiuming Pan, Yu Gong, and Philip S. Yu. 2019. Multi-Modal Generative Adversarial Network for Short Product Title Generation in Mobile E-Commerce. In *Proc. of NAACL-HLT*. 64–72.

[118] Guido Zuccon and Leif Azzopardi. 2010. Using the Quantum Probability Ranking Principle to Rank Interdependent Documents. In *Proceedings of ECIR*. 357–369.