Adversarial Web Search

Carlos Castillo
Yahoo! Research
Barcelona 08018
Catalunya-Spain
chato@yahoo-inc.com

Brian D. Davison
Lehigh University
Bethlehem, PA 18015
USA
davison@cse.lehigh.edu

Full text available at: http://dx.doi.org/10.1561/1500000021
Editorial Scope

Foundations and Trends® in Information Retrieval will publish survey and tutorial articles in the following topics:

- Applications of IR
- Architectures for IR
- Collaborative filtering and recommender systems
- Cross-lingual and multilingual IR
- Distributed IR and federated search
- Evaluation issues and test collections for IR
- Formal models and language models for IR
- IR on mobile platforms
- Indexing and retrieval of structured documents
- Information categorization and clustering
- Information extraction
- Information filtering and routing
- Metasearch, rank aggregation and data fusion
- Natural language processing for IR
- Performance issues for IR systems, including algorithms, data structures, optimization techniques, and scalability
- Question answering
- Summarization of single documents, multiple documents, and corpora
- Text mining
- Topic detection and tracking
- Usability, interactivity, and visualization issues in IR
- User modelling and user studies for IR
- Web search

Information for Librarians
Foundations and Trends® in Information Retrieval, 2010, Volume 4, 5 issues. ISSN paper version 1554-0669. ISSN online version 1554-0677. Also available as a combined paper and online subscription.
Adversarial Web Search

Carlos Castillo\(^1\) and Brian D. Davison\(^2\)

\(^1\) Yahoo! Research, Diagonal 177, 8th Floor, Barcelona 08018, Catalunya-Spain, chato@yahoo-inc.com
\(^2\) Lehigh University, 19 Memorial Drive West, Bethlehem, PA 18015, USA, davison@cse.lehigh.edu

Abstract

Web search engines have become indispensable tools for finding content. As the popularity of the Web has increased, the efforts to exploit the Web for commercial, social, or political advantage have grown, making it harder for search engines to discriminate between truthful signals of content quality and deceptive attempts to game search engines’ rankings. This problem is further complicated by the open nature of the Web, which allows anyone to write and publish anything, and by the fact that search engines must analyze ever-growing numbers of Web pages. Moreover, increasing expectations of users, who over time rely on Web search for information needs related to more aspects of their lives, further deepen the need for search engines to develop effective counter-measures against deception.

In this monograph, we consider the effects of the adversarial relationship between search systems and those who wish to manipulate them, a field known as “Adversarial Information Retrieval”. We show that search engine spammers create false content and misleading links to lure unsuspecting visitors to pages filled with advertisements or malware. We also examine work over the past decade or so that aims to
discover such spamming activities to get spam pages removed or their effect on the quality of the results reduced.

Research in Adversarial Information Retrieval has been evolving over time, and currently continues both in traditional areas (e.g., link spam) and newer areas, such as click fraud and spam in social media, demonstrating that this conflict is far from over.
# Contents

1 Introduction 1

1.1 Search Engine Spam 2

1.2 Activists, Marketers, Optimizers, and Spammers 3

1.3 The Battleground for Search Engine Rankings 5

1.4 Previous Surveys and Taxonomies 6

1.5 This Survey 7

2 Overview of Search Engine Spam Detection 9

2.1 Editorial Assessment of Spam 9

2.2 Feature Extraction 12

2.3 Learning Schemes 16

2.4 Evaluation 19

2.5 Conclusions 22

3 Dealing with Content Spam and Plagiarized Content 23

3.1 Background 24

3.2 Types of Content Spamming 27

3.3 Content Spam Detection Methods 27

3.4 Malicious Mirroring and Near-Duplicates 30

3.5 Cloaking and Redirection 31

3.6 E-mail Spam Detection 35

3.7 Conclusions 35
4 Curbing Nepotistic Linking

4.1 Link-Based Ranking

4.2 Link Bombs

4.3 Link Farms

4.4 Link Farm Detection

4.5 Beyond Detection

4.6 Combining Links and Text

4.7 Conclusions

5 Propagating Trust and Distrust

5.1 Trust as a Directed Graph

5.2 Positive and Negative Trust

5.3 Propagating Trust: TrustRank and Variants

5.4 Propagating Distrust: BadRank and Variants

5.5 Considering In-Links as well as Out-Links

5.6 Considering Authorship as well as Contents

5.7 Propagating Trust in Other Settings

5.8 Utilizing Trust

5.9 Conclusions

6 Detecting Spam in Usage Data

6.1 Usage Analysis for Ranking

6.2 Spamming Usage Signals

6.3 Usage Analysis to Detect Spam

6.4 Conclusions

7 Fighting Spam in User-Generated Content

7.1 User-Generated Content Platforms

7.2 Splogs

7.3 Publicly-Writable Pages

7.4 Social Networks and Social Media Sites

7.5 Conclusions
# Discussion

8.1 The (Ongoing) Struggle Between Search Engines and Spammers 85

8.2 Outlook 88

8.3 Research Resources 89

8.4 Conclusions 92

## Acknowledgments 93

## References 95
Introduction

Information Retrieval (IR) is a branch of computer science that deals with tasks such as gathering, indexing, filtering, retrieving, and ranking content from a large collection of information-bearing items. It is a field of study that is over 40 years old, and started with the goal of helping users locate information items in carefully curated collections, such as the ones available in libraries. In the mid-1990s, the emergence of the World Wide Web created new research opportunities and challenges for information retrieval. The Web as a whole is larger, less coherent, more distributed and more rapidly changing than the previous document collections in which IR methods were developed [9].

From the perspective of an information retrieval system such as a search engine, the Web is a mixture of two types of content: the “closed Web” and the “open Web” [37]. The closed Web comprises a small number of reputable, high-quality, carefully maintained collections which a search engine can fully trust. The “open Web”, on the other hand, includes the vast majority of Web pages, and in which document quality cannot be taken for granted. The openness of the Web has been the key to its rapid growth and success, but the same openness is the most challenging aspect when designing effective Web-scale information retrieval systems.
Introduction

Adversarial Information Retrieval addresses the same tasks as Information Retrieval: gathering, indexing, filtering, retrieving, and ranking information, with the difference that it performs these tasks in collections wherein a subset has been manipulated maliciously \[73\]. On the Web, the predominant form of such manipulation is “search engine spamming” (also known as spamdexing or Web spam). Search engine spamming is the malicious attempt to influence the outcome of ranking algorithms, usually aimed at getting an undeservedly high ranking for one or more Web pages \[92\].

Among the specific topics related to Adversarial Information Retrieval on the Web, we find the following. First, there are several forms of general Web spam including link spam, content spam, cloaking, etc. Second, there are specialized forms of Web spam for particular subsets of the Web, including for instance blog spam (splogs), opinion spam, comment spam, referrer spam, etc. Third, there are ways in which a content publisher may attempt to deceive a Web advertiser or advertiser broker/intermediary, including search spam and click spam. Fourth, there are other areas in which the interests of the designers of different Web systems collide, such as in the reverse engineering of ranking methods, the design of content filters for ads or for Web pages, or the development of undetectable automatic crawlers, to name a few.

1.1 Search Engine Spam

The Adversarial IR topic that has received the most attention has been search engine spam, described by Fetterly et al. as “Web pages that hold no actual informational value, but are created to lure Web searchers to sites that they would otherwise not visit” \[74\].

Search engines have become indispensable tools for most users \[17\]. Web spammers try to deceive search engines into showing a lower-quality result with a high ranking. They exploit, and as a result, weaken, the trust relationship between users and search engines \[92\], and may damage the search engines’ reputation. They also make the search engine incur extra costs when dealing with documents that have little or no relevance for its users; these include network costs for downloading them, disk costs for storing them, and processing costs for
indexing them. Thus, the costs of Web spam are felt both by end-users and those providing a service to them.

Ntoulas et al. [182] measured Web spam across top-level domains (TLDs) by randomly sampling pages from each TLD in a large-scale Web search engine, and then labeling those pages manually. In their samples, 70% of the pages in the .biz domain, 35% of the pages in .us and 20% of the pages in .com were spam. These are uniform random samples, while the top results in search engines are much more likely to be spam as they are the first target of spammers. In a separate study, Eiron et al. [69] ranked 100 million pages using PageRank and found that 11 out of the top 20 achieved such high ranking through link manipulation.

Ignoring Web spam is not an option for search engines. According to Henzinger et al. [98], “Spamming has become so prevalent that every commercial search engine has had to take measures to identify and remove spam. Without such measures, the quality of the rankings suffers severely.” In other words, on the “open Web”, a naïve application of ranking methods is no longer an option.

### 1.2 Activists, Marketers, Optimizers, and Spammers

The existence of Web spam pages can be seen as a natural consequence of the dominant role of search engines as mediators in information seeking processes [85]. User studies show that search engine users only scan and click the top few results for any given search [87], which means that Web page exposure and visitor traffic are directly correlated with search engine placement. Those who seek visibility need to have pages in the top positions in search engine results pages, and thus have an incentive to try to distort the ranking method.

There are many reasons for seeking visibility on the Web. Some people (activists) spam search engines to further a political message or to help a non-profit achieve its end. This is the case of most link bombs (perhaps better known as Google bombs) that spam a particular term or phrase to link it to a particular Web page. A memorable example of this manipulation is the one that affected the query “miserable failure”, which during the 2004 presidential election, returned the home page of...
George W. Bush as the first result in several Web search engines. This was the result of a coordinated effort by bloggers and Web page authors around the world. We discuss link bombing further in Section 4.

Most search engine spam, however, is created for financial gain. There is a strong economic incentive to find ways to drive traffic to Web sites, as more traffic often translates to more revenue [231]. Singhal [212] estimated the amount of money that typical spammers expected to receive in 2005: a few US dollars per sale for affiliate programs on Amazon or E-Bay, around 6 USD per sale of Viagra, and around 20–40 USD per new member of pornographic sites. Given the small per-sale commissions and the low response rates, a spammer needs to collect millions of page views to remain profitable. Further, some spam pages exist to promote or even install malware [68, 192, 193].

The incentive to drive traffic to Web sites, both for legitimate and illegitimate purposes, has created a whole industry around search engines. The objective of Search Engine Marketing (SEM) is to assist marketers in making their Web content visible to users via a search engine. SEM activities are divided by the two principal kinds of information displayed on a search results page: the editorial content and the advertising (or “sponsored search”).

Advertising on search engines today is also a ranking process, involving bidding for keywords to match to user queries, the design of the ads themselves, and the design of the “landing pages” to which users are taken after clicking on the ads. An advertiser’s goal in sponsored search is to attract more paid traffic that “converts” (i.e., buys a product or service, or performs some other action desired by the advertiser), within a given advertising budget.

Sponsored search efforts are fairly self-regulated. First, marketers have to pay the search engine for each click on the ads. Second, the marketer does not simply want to attract traffic to his Web site, but to attract traffic that leads to conversions. Thus, it is in his best interest to bid for keywords that represent the actual contents of his Web site.

\[\text{\textsuperscript{1}}\text{Some practitioners define SEM more narrowly, focusing on the sponsored search side, but from a business perspective, all of these efforts fall under marketing.}\]
1.3 The Battleground for Search Engine Rankings

Also ad market designers are careful to design systems that provide incentives for advertisers to bid truthfully.

The objective of Search Engine Optimization (SEO), on the other hand, is to make the pages of a certain Web site rank higher in the editorial side of search engines, in order to attract more unpaid or organic traffic to the target Web site.

The efforts of a search engine optimizer, in contrast, are not self-regulating, and in some cases can significantly disrupt search engines, if counter-measures are not taken. For this reason, search engines threaten SEOs that have become spammers with penalties, which may include the demotion or removal from the index of pages that use deceptive practices. The penalties that search engines apply are well known by the SEO community. Boundaries are, of course, fuzzy, as all search engines seem to allow some degree of search engine optimization.

Moran and Hunt [169] advise Web site owners on how to tell search engine spammers from SEOs. A search engine spammer tends to (i) offer a guarantee of top rankings, which no reputable firm can do as there are many variables outside their control; (ii) propose minimal changes to the pages, which indicate that they are likely to create a link farm (described in Section 4.3) instead of actually modifying the way the content is presented to users and search engines; and (iii) suggest to use server-level cloaking (described in Section 3.5) or other modifications whose typical purpose is to spam.

1.3 The Battleground for Search Engine Rankings

In general, search engine results are ranked using a combination of two factors: the relevance of the pages to the query, and the authoritativeness of the pages themselves, irrespective of the query. These two aspects are sometimes named respectively dynamic ranking and static ranking, and both have been the subject of extensive studies from the IR community (and discussed in IR textbooks [13, 58, 154]).

Some search engine spammers may be assumed to be knowledgeable about Web information retrieval methods used for ranking pages. Nevertheless, when spammers try to manipulate the rankings of a search engine, they do not know the details about the ranking methods used.
by the search engine; for instance they do not know which are the specific features used for computing the ranking. Under those conditions, their best strategy is simply to try to game any signal believed to be used for ranking.

In the early days of the Web, search engine spammers manipulated mainly the contents and URLs of the pages, automatically generating millions of pages, including incorporating repetitions or variants of certain keywords in which the spammer was interested. Next, as search engines began to use link-based signals [33, 34, 122, 183], spammers started to create pages interlinked deceptively to generate misleading link-based ranking signals.

As the search engines adapted to the presence of Web spam by using more sophisticated methods, including the usage of machine-learning-based ranking for Web pages [201], more elements of the pages were taken into consideration which pushed spammers to become more sophisticated. Next, the possibility of adding comments to forums and the existence of other world-writable pages such as wikis presented new opportunities for spammers as they allowed the insertion of arbitrary links into legitimate pages.

Recently search engines have devised other ways of exploiting the “wisdom of crowds”, e.g., through usage data to rank pages, but search engine spammers can also pose as members of the crowds and disrupt rankings as long as they are not detected. Web spam has been evolving over the years, and will continue to evolve to reflect changes in ranking methods used by popular services.

Thus, there are a variety of useful signals for ranking and each of them represents an opportunity for spammers, and in Sections 3–7 we will highlight how spammers have taken advantage of these opportunities to manipulate valuable ranking signals and what work has been done to detect such manipulation.

1.4 Previous Surveys and Taxonomies

In 2001, Perkins [189] published one of the earliest taxonomies of Web spam. This taxonomy included content spam, link spam, and cloaking. It also suggested a test for telling spam from non-spam: Spam is "any
attempt to deceive a search engine’s relevancy algorithm”, non-spam is “anything that would still be done if search engines did not exist, or anything that a search engine has given written permission to do.”

In 2005, Gyöngyi and Garcia-Molina [93] proposed a different taxonomy. This taxonomy stressed the difference between boosting techniques and hiding techniques. Boosting techniques are directly aimed at promoting a page or a set of pages by manipulating their contents or links. Hiding techniques, instead, are used by spammers to “cover their tracks”, thus preventing the discovery of their boosting techniques.

In 2007, a brief overview of Adversarial IR by Fetterly [73] appeared in ACM Computing Reviews. It included a general description of the field, and references to key articles, data sources, and books related to the subject. In the same year Heymann et al. [101] published a survey focused on social media sites, stating that in the case of social media sites, a preventive approach was possible, in addition to detection- and demotion-based approaches. Prevention is possible because in social media sites there is more control over what users can do; for example, CAPTCHAs can be incorporated to prevent automated actions, the rate at which users post content can be limited, and disruptive users can be detected and banned.

Additionally, several Ph.D. and M.Sc. theses have included elements related to Web spam. A partial list of them includes theas in the areas of link spam [95, 149, 160, 208], splogs and spam in blogs [124, 166], content spam [180], Web spam systems in general [45, 232, 236, 251], and search engine optimization [123].

We have left out the closely related subject of e-mail spam. While some methods overlap, particularly in the case of content-based Web-spam detection (which we discuss in Section 3.6), there are substantial differences between the two areas. For a survey on e-mail spam, see, e.g., Cormack [55].

1.5 This Survey

In this survey we have tried to be relatively inclusive; this is reflected in citations to about 250 publications, which we consider large for a survey on a young sub-field of study. We also intended to appeal to a
Introduction

A wide audience including developers and practitioners. For this reason, we have chosen to present general descriptions of Web spam techniques and counter-measures, and to be selective with the details.

The rest of this monograph is organized as follows:

Section 2 describes general systems for detecting search engine spam, including the choice of a machine learning method, the feature design, the creation of a training set, and evaluation methodologies.

Section 3 describes content-based spam techniques and how to detect them, as well as malicious mirroring, which is a form of plagiarism for spam purposes.

Section 4 describes link-based spam techniques and how to detect them, and covers topics such as link alliances and nepotistic linking.

Section 5 describes methods for propagating trust and distrust on the Web, which can be used for demoting spam pages.

Section 6 describes click fraud and other ways of distorting Web usage data, including Web search logs; it also deals with the subject of using search logs as part of Web spam detection systems.

Section 7 describes ways of spamming social media sites and user-generated content in general.

Finally, the discussion in Section 8 includes future research directions and links to research resources.
References


References


References


References


References


References


References


References


References


Full text available at: http://dx.doi.org/10.1561/1500000021


References


References


References


References