SIGIR WORKSHOP REPORT

Adversarial Information Retrieval on the Web (AIRWeb 2006)

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

Marc Najork
Microsoft Research, USA
najork@microsoft.com

Tim Converse
Yahoo! Search, USA
tconvers@yahoo-inc.com

1 Introduction

The attraction of hundreds of millions of web searches per day provides significant incentive for many content providers to do whatever is necessary to rank highly in search engine results, while search engine providers want to provide the most accurate results. The conflicting goals of search and content providers are adversarial, and the use of techniques that push rankings higher than they belong is often called search engine spam. Such methods typically include textual as well as link-based techniques, or their combination.

AIRWeb 2006, the Second International Workshop on Adversarial Information Retrieval on the Web, provided a focused venue for both mature and early-stage work in web-based adversarial IR. This workshop brought together researchers and practitioners concerned with the on-going efforts in adversarial information retrieval on the Web, and built on last year's successful meeting in Chiba, Japan as part of WWW2005. The workshop solicited technical papers on any aspect of adversarial information retrieval on the Web, including, but not limited to:

- search engine spam and optimization,
- crawling the web without detection,
- link-bombing (a.k.a. Google-bombing),
- comment spam, referrer spam,
- blog spam (splogs),
- malicious tagging,
- reverse engineering of ranking algorithms,
- advertisement blocking, and
- web content filtering.

Papers addressing higher-level concerns (e.g., whether 'open' algorithms can succeed in an adversarial environment, whether permanent solutions are possible, etc.) were also welcome.

Authors were invited to submit papers and synopses in PDF format. We encouraged submissions presenting novel ideas and work in progress, as well as more mature work. Submissions were reviewed by a program committee of eighteen search experts on relevance, significance, originality, clarity, and technical merit. Out of the thirteen submissions to this year's workshop, a total of six peer-reviewed papers were presented—four research presentations and two synopses of work in progress, conveying the latest results in adversarial web IR.

In addition, the workshop included an invited talk on sponsored search by Jan Pedersen of Yahoo! and a panel session with experts on blog spam, including: Tim Converse (Yahoo! Search), Dennis Fetterly (Microsoft

Research), Natalie Glance (Nielsen BuzzMetrics), Jeremy Hylton (Google), Greg Linden (Findory) and Paul Querna (Ask.com).

The complete workshop program, including full papers and presentations are available online at `http://airweb.cse.lehigh.edu/2006/`.

2 Presentations

The day was divided into two technical paper sessions, an invited talk, a panel session on blog spam, and a final discussion period. We summarize each below.

2.1 Morning Paper Session

Ricardo Baeza-Yates led the morning talks with his presentation on "Link-Based Characterization and Detection of Web Spam," while stressing that the presenting author is the one who knows least. Ricardo motivated his talk with a number of spam examples, including a pseudo-search engine that returns constant results and link farms. Noting Fetterly et al.'s 2004 statement that "in a number of these distributions, outlier values are associated with web spam", he presented the spam detection accuracy of a series of decision trees based on increasingly complex features based on statistical measures of pages in a manually-labeled UK dataset. Simple measures included indegree, reciprocity, maximum PageRank of page in host; more sophisticated measures considered variations of TrustRank and Truncated PageRank; the final approach probabilistically counted "supporters" at different distances. While combinations of features generally (but not always) improved performance, the detection accuracy was still only about 80%, causing Ricardo to note that it was not a magic bullet.

Tanguy Urvoy presented a novel approach to the problem of spam detection with his talk on "Tracking Web Spam with Hidden Style Similarity." Rather than looking at links or content, he examined the formatting of the page source. His approach was to focus on what most would consider noise by simply stripping out all alphanumeric characters, and to detect if two or more HTML pages were generated by the same tools and templates. Clusters were then created of highly-templatic content sources for easy detection.

Rashmi Raj's synopsis, "Web Spam Detection with Anti-Trust Rank" focused on the propagation of untrustworthiness backward from a seed set of untrusted pages. This approach was analogous (although inverted) to Gyongyi et al.'s TrustRank.

2.2 Invited Talk

Jan Pedersen, chief scientist of Yahoo! Search and Marketplace, delivered his invited talk on "Sponsored Search: Theory and Practice." After reminding us why sponsored search is important (a successful, large market!), Jan described the history of sponsored search (starting with GoTo.com and ending with the improvements introduced by Google), including that GoTo's origin was in part a response to the practice of search engine optimization (with a transparent ranking function). Jan then presented an analysis framework based on auction theory to precisely define the various kinds of ranking mechanisms used. He concluded his presentation with a discussion of practical issues such as keyword matching concerns, estimating click rates, and click fraud and traffic quality, and revenue maximization.

2.3 Afternoon Paper Session

After lunch, Bernard (Jim) Jansen continued the sponsored search theme with his synopsis on "Adversarial Information Retrieval Aspects of Sponsored Search." He pointed out that there has not yet been much academic work on sponsored search, even though it contains a significant information retrieval component.

He also noted the adversarial problem of click fraud within sponsored search. Jim described how click fraud is performed and gave suggestions on how to combat and prevent it.

Károly Csalogány gave an admirable first presentation in English, entitled "Link-Based Similarity Search to Fight Web Spam." This work introduced the use of link-based similarity measures (e.g., co-citation, SimRank) to detect spam, and compared it to existing trust and distrust propagation techniques on two country-specific datasets. Similarity-based methods were shown to have better performance.

Kumar Chellapilla presented his work on "Improving Cloaking Detection using Search Query Popularity and Monetizability." After motivating the problem of cloaking with a number of striking examples (such as a frame that obscured content beneath), he presented his hypothesis that the more monetizable a query is, the more likely it is to be spammed. This was tested by collecting the 5,000 most popular queries from query logs, and the 5,000 most monetizable queries (those that generated the most revenue from sponsored ads on a single day). By applying an extension to Wu and Davison's (2005, 2006) technique for cloaking detection, Chellapilla and Chickering found 9.7% cloaking in monetizable queries, and 6.0% cloaking in popular queries.

2.4 Expert Panel Session

This year's panel consisted of experts in blog spam, including Tim Converse (Yahoo! Search), Dennis Fetterly (Microsoft Research), Natalie Glance (Nielsen BuzzMetrics), Jeremy Hylton (Google), Greg Linden (Findory), and Paul Querna (Ask.com). During the first part of the session, the panelists introduced themselves and outlined their perspectives on blog spam.

Natalie Glance is one of the creators of BlogPulse, a blog search and analytics website backed by Nielsen BuzzMetrics. She stated that blog spam, like other web spam, is driven by search engine optimizers, and pointed out that one of reasons why blogs are attractive targets for spammers is that they provide effectively free hosting of sponsored ads and link farms. While not targeting blog search, blog spam will pollute blog search results. She also noted that blog spam distorts statistical measures (e.g. trend graphs), which is particularly annoying to blog analytics companies such as BuzzMetrics.

Jeremy Hylton works on blog search for Google, and started by elaborating on the temporal dimension of blogs. Postings are dated, which allows search engines to rank them in reverse temporal order (giving high rank to recent postings). However, very recent posts by definition did not have a chance to attract links, making traditional link-based ranking algorithms (such as PageRank) ill-suited for the blog domain. Moreover, time-based ranking schemes can be overwhelmed by the sheer volume of content—spammers can automatically generate large streams of postings. Jeremy explained that Google's blog search doesn't crawl blogs, but instead uses RSS and Atom feeds (which avoids the problem of cloaked pages). In later discussion, Hylton additionally noted that there is little comment spam in feeds (as opposed to web pages). He also claimed that blog search does not have to worry about navigational queries. Finally, Jeremy presented some graphs showing blog spam over time. According to these graphs, about 10-20% of all postings are spam (with occasional spikes to 50%), while 2-10% of all blogs consist entirely of spam (spam blogs or "splogs").

Paul Querna supports Bloglines, an Ask.com product. A few months ago, it launched a blog search service. By indexing and monitoring what users subscribe to and read on Bloglines, Paul reports that Ask is able to avoid spam blogs since it is much harder for automated tools to impersonate a human reader than to generate bogus content. This fact, coupled with the observation that humans do not like spam and generally will not read, it makes it possible to leverage the human readership of RSS aggregators in the job of spam detection. In response to a later question, Paul confirmed that they have seen fake users in the system.

Greg Linden operates Findory.com, which uses personalization to recommend blog articles based on what you read. He cited Technorati and Bloglines statistics that suggested that most blogs (99.8%) have no regular readers. He argued that the winner take all effect in web search (where top-ranked results garner most of the traffic) is a driver for spam, and that personalization seems to be a solution.

Tim Converse heads the algorithmic anti-spam group at Yahoo! Search. He pointed out that publicly writable pages allow for the violation of the propagation-of-trust along links assumption that is at the basis of the various link-based ranking algorithms used by search engines. One attempt to fight comment spam

is the *nofollow* link attribute for HTML anchors, introduced more than a year ago, which indicates that the link should not be trusted. According to Tim, about 1% of links use *nofollow* (corresponding to billions of links), so the attribute is in all likelihood an improvement. Tim is more concerned with the newer tactic of blogs composed entirely of spam postings (splogs), which are used for link creation, not as a destination. He finds that RSS and syndication provides a cheap and powerful framework for spammers to acquire human-authored (and often high-quality) content, which they can then repurpose to generate spam web pages (or spam postings). He notes that there is a spectrum from intelligent aggregation and spam-purposed text scraping, weaving and stitching. Finally, he predicts increasing cooperation and authentication between search engines and content providers to address the spam problem.

Dennis Fetterly of Microsoft Research continued the concerns about repurposed content in splogs and in general, giving a number of examples of the same content (including both legitimate news sites as well as link farm spam). He shared the results of manual labeling more than 2500 blogs: 13.7% were spam, and noted that 39% of spam blog pages were from four popular blog hosting sites.

The second part of the panel consisted of questions by the audience to the panelists. Asked to break down blog spam, the panelists agreed that most blog spam is aimed at link generation (to game PageRank-style algorithms), followed by direct revenue generation through embedded advertisements. In the opinion of the panelists, human-interactive proofs (a.k.a. "Captchas") are a highly effective way to reduce comment spam; the main argument against them being the accessibility issues of image-based captchas. Even very simple captchas (such as asking users to answer the question "how much is two plus two?" before admitting a post) are currently effective. While there was agreement that it would be good to have a "clearinghouse" for blog spam, the panelists believed that the competitive nature of the search industry makes it difficult to establish such an institution.

We did see some differences in panelist opinions, typically reflecting the technology used. For example, the two panelists using collaborative recommendations (Findory's Linden and Bloglines's Querna) agreed that there is never a good reason to show a splog as the result of a blog search. Those representing the biggest engines (Yahoo's Converse and Google's Hylton), said they would show it, but would downgrade the spam.

2.5 Discussion

Marc Najork provided some concluding remarks and found wide support for continuing the workshop series, and general agreement to co-locate with WWW in 2007. *Authors' note: a workshop proposal for WWW in 2007 is underway.*

3 Acknowledgments

We extend our sincere thanks to SIGIR, to the authors and presenters, to the expert panelists and invited speaker, and to the members of the program committee for their contributions to the material that formed an outstanding workshop.