WORKSHOP REPORT

Adversarial Information Retrieval on the Web (AIRWeb 2007)

Carlos Castillo
Yahoo! Research
chato@yahoo-inc.com

Kumar Chellapilla
Microsoft Live Labs
kumarc@microsoft.com

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

1 Introduction

The ubiquitous use of search engines to discover and access Web content shows clearly the success of information retrieval algorithms. However, unlike controlled collections, the vast majority of Web pages lack an authority asserting their quality. This openness of the Web has been the key to its rapid growth and success, but this openness is also a major source of new adversarial challenges for information retrieval methods.

Adversarial Information Retrieval addresses tasks such as gathering, indexing, filtering, retrieving and ranking information from collections wherein a subset has been manipulated maliciously. On the Web, the predominant form of such manipulation is "search engine spamming" or spamdexing, i.e., malicious attempts to influence the outcome of ranking algorithms, aimed at getting an undeserved high ranking for some items in the collection. There is an economic incentive to rank higher in search engines, considering that a good ranking on them is strongly correlated with more traffic, which often translates to more revenue.

AIRWeb 2007, the Third International Workshop on Adversarial Information Retrieval on the Web, provided a focused venue for both well developed and early-stage research work in web-based adversarial IR. This workshop brought together researchers and practitioners working on topics relating to adversarial information retrieval on the Web, and built on two successful prior meetings in Chiba, Japan as part of WWW2005, and Seattle, USA as part of SIGIR2006. The workshop solicited technical papers on any aspect of adversarial information retrieval on the Web, including, but not limited to:

- Link spam: nepotistic linking, collusion, link farms, link exchanges and link bombing
- Content spam: keyword stuffing, phrase stitching, and other techniques for generating synthetic text
- Cloaking: sending different content to a search engine than to regular visitors of a web site, which is often used in combination with other spamming techniques
- Comment spam in legitimate sites: in blogs, forums, wikis, etc
- Spam-oriented blogging: splogs, spings, etc
- Click fraud detection: including forging clicks for profit, or to deplete a competitor's advertising funds
- Reverse engineering of ranking algorithms
- Web content filtering: as used by governments, corporations or parents to restrict access to inappropriate content
- Advertisement blocking: developing software for blocking advertisements during browsing
- Stealth crawling: crawling the Web while avoiding detection
- Malicious tagging: for injecting keywords or for self-promotion in general.

Papers addressing higher-level concerns, such as whether "open" algorithms can succeed in an adversarial environment, whether permanent solutions are possible, how the problem has evolved over years, what are the differences between "black-hat", "white-hat", and "gray-hat" techniques, where is the line between search engine optimization and spamming, 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 thirty-three search experts on relevance, significance, originality, clarity, and technical merit. Out of the twenty submissions to this year's workshop, a total of thirteen peer-reviewed papers were presented—ten research presentations and three synopses of work in progress, conveying the latest results in adversarial web IR.

In addition, the workshop included a web spam challenge for testing web spam detection systems. This challenge was supported by the EU Network of Excellence PASCAL Challenge Program, and by the DELIS EU-FET research project. Participation was open to all. The WEBSPAM-UK2006 collection [3] comprising a large set of web pages, a web graph, and human-provided labels for hosts was used as the dataset. A sample set of content and link features was also provided to make it easy for participating teams. Participating teams submitted predictions (normal/spam) for unlabeled hosts in the collection. Systems were evaluated based on the F1 statistic and AUC (Area under the ROC curve) metric. Participation in the challenge did not require a paper submission. However, participants were encouraged to submit research articles describing their systems. Nine web spam detection system entries were received from six different teams.

The complete workshop program, including full papers and presentations are available online at http://airweb.cse.lehigh.edu/2007/.

2 Presentations

The day was divided into three technical paper sessions and a web spam challenge session. The three technical paper sessions focused on: Temporal and Topological Factors, Link Farms, and Tagging, P2P and Cloaking. Winners of the web spam challenge presented their systems during the web spam challenge session. We summarize each below.

2.1 Session 1: Temporal and Topological Factors

Belle Tseng started the morning session with her presentation on "Splog Detection Using Self-Similarity Analysis on Blog Temporal Dynamics" [9]. After a few quick motivating examples of splogs (a blog created with the intention of spamming), Belle gave an insight into the magnitude of the problem. Even though only 10-20% of all blogs are splogs, over 75% of new pings come from splogs, and 44% of top-100 results from the top-3 blog search engines are splogs. She presented their solution to splog detection which has three salient features: self-similarity analysis, visual characterization, and temporal feature computation. The visual characterization of blog post times using a clock-like representation was very intuitive and made it easy to separate splogs from normal blogs. Belle showed that classifiers built on temporal features alone out-performed those built on content features alone with the overall system achieving 95% precision at 94% recall on the TREC Blog Track 2006 collection.

Krysta Svore presented an improved approach to classifying web spam in her talk on "Improving Web Spam Classification using Rank-time Features" [13]. While identifying that a search results ranker is not trained to detect spam, Krysta noted that the ranker is the last stage where spam can be caught by the search engine before it shows up in search results. A spam aware ranker is well suited for detecting web spam results that might even have high relevance to the user query. Krysta highlighted that it is crucial to separate domains such that they do not get split between training and test sets, i.e., the domain used in the training set must not occur in the test set and vice versa. This ensures good generalization to unseen domains. Not doing so can produce as much as a 40% difference in precision between the trained classifier and field tests. Results on a large web spam dataset showed that using rank time features can improve classification by as much as 25% in recall at a set precision.

Qingqing Gan presented a summary of a two-stage approach to improving spam classifiers [7]. The first stage targets recall by preclassifying web pages and the second stage uses several heuristics that examine web page neighborhood (along with other features) to determine whether the page should be relabeled or processed by a second classifier.

Krysta Svore took the stage again, this time to talk about “Transductive Link Spam Detection.” She presented a novel link spam detection algorithm based on semi-supervised learning on directed graphs. Krysta pointed out that regularization is key to building a classifier that generalizes well and discrete regularization on graphs implies that it lets the classification function value change slowly over densely connected subgraphs.

2.2 Session 2: Link Farms

Ye Du started the second session with his talk on “Using Spam Farm to Boost PageRank.” He reviewed the spam farm model, wherein a spam farm is comprised of a single target page and a number of boosting pages. An optimal spam farm maximizes the PageRank score of the target page. Ye presented his improved analysis that relaxed the constant leakage assumption. He also introduced constraints such as allowing only a subset of the boosting pages to link directly to the target page. Ye presented new theoretical results on the structure of optimal link farms for these generalized cases.

Baoning Wu presented his work on “Extracting Link Spam using Biased Random Walks from Spam Seed Sets.” Baoning observed that link farms and link exchanges typically form dense cross-linking structures that can be viewed as clusters in the web graph. He presented a random walk based approach that begins from a seed web page marked as spam and explores its local graph neighborhood and extracts the link exchange/farm community. Baoning presented experimental results using manually labeled link spam data sets and random walks from a single seed domain that showed that approach achieves over 95.12% precision in extracting large link farms and 80.46% precision in extracting link exchange centroids.

Masashi Toyoda presented “A Large-Scale Study of Link Spam Detection by Graph Algorithms.” He presented graph algorithms that compute strongly connected components (SCC), maximal cliques, and expand found link farms using min-cut. He showed that on the Japanese web archive dataset, the size distribution of SCCs follows a power law with a long tail and that most of the large SCCs and cliques were link farms.

Xiaoguang Qi continued the session on link farms with his talk on “Measuring Similarity to Detect Qualified Links.” Xiaoguang highlighted that not all links on a page are qualified to make a recommendation regarding the target page. He showed several examples of spam links, navigational links, advertising links, and other irrelevant links that are effectively noise for link analysis algorithms. He then went on to present a filtering procedure that uses link classifiers built through machine learning to remove noisy links. He demonstrated that a qualified version of HITS filtered 37% of links and boosted precision by 9% over previous best results.

2.3 Session 3: Tagging, P2P and Cloaking

Georgia Koutrika opened the afternoon session with her talk on “Combating Spam in Tagging Systems.” Tag spam refers to misleading tags that are generated to increase visibility of some resources, or generated simply to confuse users. She pointed out that common causes of Tag spam are user error, malicious intent and commercial intent. Using an ideal model of a tagging system she showed that current tagging systems that rely on using the number of occurrences of a tag in a document’s postings for answering tag queries, are threatened not only by malicious users but also by user errors. Georgia recommended using trusted moderators as a countermeasure.

Debora Donato addressed decentralized reputation management in her talk on “New Metrics for Reputation Management in P2P Networks.” Decentralized reputation management is used to describe the performance and reliability of machines in file sharing peer-to-peer networks. She presented an improved approach that combined web spam style metrics with the original EigenTrust algorithm for detecting malicious peers participating in sophisticated attacks.

Continuing the peer-to-peer theme, Josiane Parreira presented a new approach to “Computing Trusted Authority Scores in Peer-to-Peer Web Search Networks.” She addressed improvements to distributed computing of PageRank scores for information units (Web pages, sites, peers, social groups, etc.) within a link- or endorsement-based graph structure. She motivated the need for recognizing that some peers are not always honest (called cheating peers). She then presented TrustJXP which is robust to attacks where peers report higher scores or permute scores for a subset of their local pages.

Kumar Chellapilla presented “A Taxonomy of JavaScript Redirection Spam.” JavaScript redirection is the most notorious of redirection techniques and is hard to detect as many of the prevalent crawlers are script-agnostic. He motivated the problem through several examples. Using a large dataset containing over 750K popular pages and 930K blog pages, he presented a taxonomy of the different types of JavaScript Redirection Spam and their prevalence on the web. He presented findings that showed that while 25% of all JavaScript redirection spam examines the referrer property, 44% of popular and 62% of blog pages use obfuscation techniques that limit the effectiveness of static analysis and static feature based systems. Based on these findings, he recommended a robust counter measure using a light weight JavaScript parser and engine.

István Bíró presented a synopsis of their web spam detection system in his talk on “Web Spam Detection via Commercial Intent Analysis.” He demonstrated that their system improved spam detection numbers by 3% on the UK2006 dataset by combining features that captured commercial intent with existing online commercial classifiers such as Online Commercial Intention (OCI) value from Microsoft adCenter Labs, Yahoo! Mindset classification, as well as metrics based on the occurrence of Google ads on the page. Their system was also submitted as an entry to the web spam challenge.

2.4 Web Spam Challenge

The final session summarized the entries and results from the Web Spam Challenge. The Web Spam Challenge received nine entries from six teams. The challenge was successful in bringing together academia and companies to produce web spam solutions for real-world data. The competition attracted three teams from academic institutions (Hungarian Academy of Sciences, University of Waterloo, and Chinese Academy of Sciences) and three teams from research labs in companies (Genie Knows, Microsoft, and France Telecom).

Gordon Cormack from the University of Waterloo adapted a stack of state-of-the-art e-mail content-based classifiers for Web spam classification. The content used for each host was its home page, plus its host name and the hostnames of its neighbors (through links). Both Dennis Fetterly from Microsoft Research, and Istvan Biro from the Hungarian Academy of Sciences presented machine learning based approaches and expanded the feature sets with novel link- and content-based features. The team from the Institute of Automation at the Chinese Academy of Sciences used under-sampling of the hosts in the training set, and demonstrated its impact in performance improving of an automatic classifier.

Tapajyoti Das presented a graph-based approach, in which some labels in the training set (both spam and nonspam) were propagated recursively through links. Finally, Pascal Filoche presented a clustering method that aggregated hosts created by Web page generation/preprocessing software.

The testing performance of the web spam detection systems, ranged from 0.67 to 0.91 under the F-Measure. The Area under the ROC curve (AUC) metric ranged from 0.80 to 0.96 between the participants. Most of the teams did well, with the majority of them achieving F-Measure scores higher than 0.8 and AUC scores above 0.93. All teams expressed that the competition was challenging as it required significant effort due to the data processing requirements but was also fun and motivating.

For more information, a detailed description of the entries in the challenge can be found at http://webspam.lip6.fr/wiki/pmwiki.php?n=Main.PhaseIResults

2.5 Discussion

Carlos Castillo provided some concluding remarks and found wide support for continuing the workshop series, and general agreement to co-locate with WWW in 2008. *Authors’ note: AIRWeb 2008 is co-located with WWW 2008 in Beijing, China.*

3 Acknowledgments

We extend our sincere thanks to WWW2007, to the authors and presenters, and to the members of the program committee for their contributions to the material that formed an outstanding workshop.

References

[1] A. Benczúr, I. Bíró, K. Csalogány and T. Sarlós, Web spam detection via commercial intent analysis. In [2], pages 89–92, May 2007,

[2] C. Castillo, K. Chellapilla, and B.D. Davison, eds. *AIRWeb '07: Proceedings of the 3rd International Workshop on Adversarial Information Retrieval on the Web*. ACM International Conference Proceedings Series, Vol. 215, ACM Press. May, 2007.

[3] C. Castillo, D. Donato, L. Becchetti, P. Boldi, M. Santini and S. Vigna, A Reference Collection for Web Spam. *ACM SIGIR Forum* 40(2):11-24, December 2006.

[4] K. Chellapilla and A. Maykov, A taxonomy of JavaScript redirection spam. In [2], pages 81-88, May 2007.

[5] D. Donato, M. Paniccia, M. Selis, C. Castillo, G. Cortese, and S. Leonardi, New metrics for reputation management in P2P networks. In [2], pages 65-72, May 2007.

[6] Y. Du, Y. Shi, and X. Zhao, Using spam farm to boost PageRank. In [2], pages 29-36, May 2007.

[7] Q. Gan and T. Suel, Improving web spam classifiers using link structure. In [2], pages 17-20, May 2007.

[8] G. Koutrika, F.A. Effendi, Z. Gyöngyi, P. Heymann, and H. Garcia-Molina, Combating spam in tagging systems. In [2], pages 57-64, May 2007.

[9] Y-R. Lin, H. Sundaram, Y. Chi, J. Tatemura, and B.L. Tseng, Splog detection using self-similarity analysis on blog temporal dynamics. In [2], pages 1-8, May 2007.

[10] J.X. Parreira, D. Donato, C. Castillo, G. Weikum, Computing trusted authority scores in peer-to-peer web search networks. In [2], pages 73-80, May 2007,

[11] X. Qi, L. Nie, and B.D. Davison, Measuring similarity to detect qualified links. In [2], pages 49–56, May 2007.

[12] H. Saito, M. Toyoda, M. Kitsuregawa, and K. Aihara, A large-scale study of link spam detection by graph algorithms. In [2], pages 45-48, May 2007.

[13] K.M. Svore, Q. Wu, C. Burges, and A. Raman, Improving web spam classification using rank-time features. In [2], pages 9-16, May 2007.

[14] B. Wu and K. Chellapilla, Extracting link spam using biased random walks from spam seed sets. In [2], pages 37-44, May 2007.

[15] D. Zhou, C. Burges, and T. Tao, Transductive link spam detection. In [2], pages 21–28, May 2007.