Fourth International Workshop on Adversarial Information Retrieval on the Web (AIRWeb 2008)

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

Kumar Chellapilla
Microsoft Live Labs
Redmond, WA, USA
kumarc@microsoft.com

Dennis Fetterly
Microsoft Research
Mountain View, CA, USA
fetterly@microsoft.com

ABSTRACT

Adversarial IR in general, and search engine spam, in particular, are engaging research topics with a real-world impact for Web users, advertisers and publishers. The AIRWeb workshop will bring researchers and practitioners in these areas together, to present and discuss state-of-the-art techniques as well as real-world experiences. Given the continued growth in search engine spam creation and detection efforts, we expect interest in this AIRWeb to surpass that of the previous three editions of the workshop (held jointly with WWW 2005, SIGIR 2006, and WWW 2007 respectively).

Categories and Subject Descriptors

H.3.5 [**Information Storage and Retrieval**]: Online Information Services

General Terms

Documentation

Keywords

Adversarial information retrieval, web spam, spamdexing, search engine spam

1. INTRODUCTION

Before the advent of the World Wide Web, information retrieval algorithms were developed for relatively small and coherent document collections such as newspaper articles or book catalogs in a library. In comparison to these collections, the Web is massive, much less coherent, changes more rapidly, and is spread over geographically distributed computers [1]. Scaling information retrieval algorithms to the World Wide Web is a challenging task. Success to date is depicted by the ubiquitous use of search engines to access Internet content.

From the point of view of a search engine, the Web is a mix of two types of content: the "closed Web" and the "open Web" [2]. The closed web comprises a few high-quality controlled collections which a search engine can fully trust. The "open Web," on the other hand, includes the vast majority of Web pages, which lack an authority asserting their quality. The openness of the Web has been the key to its rapid growth and success. However, this openness is also a major source of new challenges for information retrieval methods.

Search engine spam is not a new problem; it has been an important issue for commercial providers for a number of years, and is not likely to be solved in the near future. Web spam damages search engine reputation. It exploits and as a result weakens the trust relationship between users and search engines [6]. According to Henzinger *et al.* [7], "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." On the "open web" a naive application of ranking methods in no longer an option. For instance, PageRank [8] in its pure form is very susceptible to spam: the authors of [4] ranked 100 million pages using PageRank and found that 11 out of the top 20 were pornographic and achieved such high ranking through link manipulation.

Adversarial information retrieval is a research area in which several things remain to be discovered. Sahami *et al.* [9] have noted that "Adversarial classification is an area in which precious little work has been done, but effective methods can provide large gains." Also, adversarial IR problems can be approached from many different perspectives, including information retrieval, machine learning and game theory.

2. WORKSHOP TOPICS

Adversarial Information Retrieval addresses tasks such as gathering, indexing, filtering, retrieving and ranking information from collections wherein a subset has been manipulated maliciously [5]. 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 [10].

As in previous years, automatic detection of search engine spam is expected to be the dominant theme of this workshop. Three basic forms of web spam are included:

- **Link spam**
- **Content spam**
- **Cloaking**

WWW 2008, April 21–25, 2008, Beijing, China.
ACM 978-1-60558-085-2/08/04.

Several other adversarial IR topics that we welcome include:

- **Blog spam filtering**
- **Click fraud detection**
- **Reverse engineering of ranking algorithms**
- **Web content filtering**
- **Advertisement blocking**
- **Stealth crawling**

3. WEB SPAM CHALLENGE

In 2007, we introduced a novel element: the Web spam challenge. We released a reference collection for Web Spam Detection that comprises Web pages, a Web graph, and labels for a subset of the pages. Web pages in this collection were labeled as "normal" or "spam" by humans [3]. Using this data set, the challenge was to predict which pages in the unlabeled part of the data are spam and which are normal. For 2008, we released an updated reference collection covering a significantly increased number of hosts. We also encouraged authors submitting papers on search engine spam to test their systems on the updated reference collection.

We ask that participating researchers submit predictions (normal/spam) for all unlabeled elements in the collection. Predictions will be evaluated on a part of the collection for which human-provided labels will be held for testing. Results will be announced at the AIRWeb 2008 workshop.

The Web spam challenge serves a dual purpose: it allows the comparison of different systems, which has not been possible in the past for lack of a reference collection; and it stimulates research on this area given its competitive nature.

4. WORKSHOP ORGANIZATION

The proceedings of the workshop will be published online in the ACM Digital Library, as well as distributed at the workshop. The workshop program has not been finalized at the time of this writing. Once finalized, the program will be available from the AIRWeb website:
`http://airweb.cse.lehigh.edu/2008/`

4.1 Program Committee

We appreciate the service of the following researchers as Program Committee members of the workshop:

- Einat Amitay – IBM
- András Benczúr – Hungarian Academy of Sciences
- James Caverlee – Texas A&M University
- Paul-Alexandru Chirita – Adobe
- Gordon Cormack – University of Waterloo
- Nick Craswell – Microsoft Research
- Matt Cutts – Google
- Brian Davison – Lehigh University
- Ludovic Denoyer – University Paris 6
- Aaron D'Souza – Google
- Edel Garcia – Mi Islita.com
- Natalie Glance – Nielsen BuzzMetrics
- Antonio Gulli – Ask.com
- Zoltán Gyöngyi – Stanford University
- Monika Henzinger – Google
- Pranam Kolari – Yahoo! Applied Research
- Mark Manasse – Microsoft Research
- Marc Najork – Microsoft Research
- Alexandros Ntoulas – Microsoft Research
- Jan Pedersen – Yahoo! Research
- Erik Selberg – Amazon.com
- Torsten Suel – Polytechnic University
- Mike Thelwall – University of Wolverhampton
- Tao Yang – Ask.com
- Baoning Wu – Snap.com

5. REFERENCES

[1] Arvind Arasu, Junghoo Cho, Hector Garcia-Molina, Andreas Paepcke, and Sriram Raghavan. Searching the web. *ACM Transactions on Internet Technology (TOIT)*, 1(1):2–43, 2001.

[2] Terrence A. Brooks. Web search: how the Web has changed information retrieval. *Information Research*, 8(3), April 2003.

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

[4] Nadav Eiron, Kevin S. Curley, and John A. Tomlin. Ranking the web frontier. In *Proceedings of the 13th international conference on World Wide Web*, pages 309–318, New York, NY, USA, 2004. ACM Press.

[5] Dennis Fetterly. Adversarial information retrieval: The manipulation of web content. *ACM Computing Reviews*, July 2007.

[6] Zoltán Gyöngyi and Hector Garcia-Molina. Spam: It's not just for inboxes anymore. *IEEE Computer Magazine*, 38(10):28–34, 2005.

[7] Monika R. Henzinger, Rajeev Motwani, and Craig Silverstein. Challenges in web search engines. *SIGIR Forum*, 36(2):11–22, 2002.

[8] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: bringing order to the Web. Technical report, Stanford Digital Library Technologies Project, 1998.

[9] Mehran Sahami, Vibhu Mittal, Shumeet Baluja, and Henry Rowley. The happy searcher: Challenges in web information retrieval. In *Trends in Artificial Intelligence, 8th Pacific Rim International Conference on Artificial Intelligence (PRICAI)*, pages 3–12, Auckland, New Zealand, August 2004. Springer.

[10] Yi-Min Wang, Ming Ma, Yuan Niu, and Hao Chen. Spam double-funnel: connecting web spammers with advertisers. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 291–300, New York, NY, USA, 2007. ACM Press.